Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants

Natalie Fecher

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

University of York
Department of Language and Linguistic Science

Submitted July 2014

Abstract

This thesis offers a thorough investigation into the effects of forensically-relevant facial concealment on speech acoustics and perception. Specifically, it explores the extent to which selected acoustic-phonetic and auditory-perceptual properties of consonants are affected when the talker is wearing 'facewear' while speaking. In this context, the term 'facewear' refers to the various types of face-concealing garments and headgear that are worn by people in common daily communication situations; for work and leisure, or as an expression of religious, social and cultural affiliation (e.g. surgical masks, motorcycle helmets, ski and cycling masks, or full-face veils such as the *niqāb*). It also denotes the face or head coverings that are typically used as deliberate (visual) disguises during the commission of crimes and in situations of public disorder (e.g. balaclavas, hooded sweatshirts, or scarves).

The present research centres on the question: does facewear influence the way that consonants are produced, transmitted, and perceived? To examine the effects of facewear on the acoustic speech signal, various intensity, spectral, and temporal properties of spoken English consonants were measured. It was found that facewear can considerably alter the acoustic-phonetic characteristics of consonants. This was likely to be the result of both deliberate and involuntary changes to the talker's speech productions, and of sound energy absorption by the facewear material. The perceptual consequences of the acoustic modifications to speech were assessed by way of a consonant identification study and a talker discrimination study. The results of these studies showed that auditory-only and auditory-visual consonant intelligibility, as well as the discrimination of unfamiliar talkers, may be greatly compromised when the observer's judgements are based on 'facewear speech'.

The findings reported in this thesis contribute to our understanding of how auditory and visual information interact during natural speech processing. Furthermore, the results have important practical implications for legal cases in which speech produced through facewear is of pivotal importance. Forensic speech scientists are therefore advised to take the possible effects of facewear on speech into account when interpreting the outcome of their acoustic and auditory analyses of evidential speech recordings, and when evaluating the reliability of earwitness testimony.

Contents

List of Tables

List of Figures

Acknowledgments

While working on this thesis, I was very fortunate to meet many wonderful people and inspiring researchers who have, in their very own ways, done a great job to help me stay on track (with my research, and otherwise). In the following, I would like to take the opportunity to express my thanks and appreciation to some of them.

First and foremost, my gratitude goes to Dr. Dominic Watt for his continued support, enthusiasm, immense knowledge and commitment. He encouraged me to pursue my own ideas, while never failing to offer me professional advice in all my decisions along the way. Dom, this thesis is as much yours as it is mine, not least because it was you who developed the research idea in the first place. I thank you for teaching me how to write fairly decent and thought-out English sentences, for not despairing of my lack of understanding of English prepositions, for your eye (and ear) for detail, for balancing my (academic) ups and downs, and for our regular exchange of articles on the latest bank robber fashion and other scientific niches and curiosities.

I would like to extend my appreciation to the members of my thesis advisory panel in York, Dr. Carmen Llamas and Prof. Paul Foulkes. Thank you both for your encouragement and constructive feedback, and helping me enrich my research. I am moreover very grateful to Prof. Peter French for his support, and for giving me the invaluable opportunity to gain practical experience with forensic phonetic casework. I also thank him and Prof. Volker Dellwo for the time and effort they put into reading and evaluating this thesis, and for the feedback and advice given during my viva.

This research would not have been possible without the opportunities provided by the Marie Curie Initial Training Network ‘BBfor2’. Being a member of this network has helped me to grow as a researcher, and the experience as a whole has been priceless. A warm thanks to Prof. David van Leeuwen and Dr. Henk van den Heuvel for the tremendous amount of work they put into organising and leading the project.

I specially thank my advisor in the BBfor2 network, Prof. David van Leeuwen. It was my great pleasure to work with you, David, during my secondment at the Centre for Language and Speech Technology at Radboud University Nijmegen. I thank you for our (not at all confusing) discussions about my consonant confusion data, and for sharing your expertise in *d-prime* and *SINFA* analysis with me (and for the scripts!).

My sincere gratitude extends to colleagues from other universities and hemispheres, and in particular to Prof. Denis Burnham for facilitating my stay at the MARCS Institute, University of Western Sydney, and to Profs. Chris Davis and Jeesun Kim, with whom I was fortunate to work for some time. Chris and Jeesun, your knowledge and commitment to the highest standards inspired and motivated me. I cannot thank you enough for your insightful comments and constructive criticisms of my research.

I moreover thank Prof. Martin Cooke, who helped me with the sound mixing for one of my studies and offered excellent advice on my acoustic speech data. Thanks for the 'babble'! I hope to meet you face-to-face some day to thank you again in person.

At the University of York, my thanks also go to Huw Llewelyn-Jones for technical support and assistance during data collection (and for being the best Wednesday circuits trainer!); to Tai Chi Minh Ralph Eastwood for his programming skills; to Omer Qadir for his help with signal processing and all things Linux (and Vim!); to Christin Kirchhübel for proofreading this thesis; to Philip Harrison and all other members of the Forensic Speech Science research group at the Department of Language and Linguistic Science; to Sven Mattys for helpful advice; and last but certainly not least, to Erica Gold for being the most considerate and supportive office mate, and for going on this lively 'BBfor2 journey' with me.

At the MARCS Institute in Sydney, thanks are due to all colleagues who helped me find my way inside and outside the institute; to Benjamin Schultz for assistance with experimental design and data analysis; and to Michael Fitzpatrick for sharing my enthusiasm for AV research (and for advising me how not to run into deadly spiders!). For encouraging and inspiring conversations I would like to thank Rahim Saeidi, Odette Scharenborg, James McQueen and Margriet Groen at Radboud University Nijmegen, as well as Charlie Frowd and Amanda Heath at the University of Central Lancashire. At the Idiap Research Institute in Martigny, I thank Laurent El Shafey for his help. At the Institute of Phonetics and Speech Processing in Munich, heartfelt thanks go to Christoph Draxler for his encouraging words and support throughout the years; to Florian Schiel for assistance with *MAUS*; and to Jonathan Harrington for introducing me to (forensic) phonetics in the first place.

In addition, I would like to acknowledge the 120+ participants in my experiments, who lent me their voices and/or ears. Thanks to you all for your time and interest!

Words cannot express how thankful I am to my loved ones and friends for being a constant source of support, patience and strength. You make me climb higher and higher (literally and figuratively)! Above all, I thank my most wonderful parents and best sister ever. Ich danke Euch aus tiefstem Herzen, dass Ihr jederzeit ein offenes Ohr für mich habt, mich bei all meinen Entscheidungen unterstützt, und immer mein Sicherheitsnetz in diesem Hochseilakt des Lebens seid.

Finally, I must thank the generous financial support provided by the Marie Curie Initial Training Network 'Bayesian Biometrics for Forensics (BBfor2)', which was funded by the Seventh Framework Programme of the European Communities (FP7-PEOPLE-ITN-2008) under grant agreement number 238803.

Author’s declaration

This is to certify that the present thesis has not previously been submitted for any degree other than Doctor of Philosophy of the University of York. It is furthermore confirmed that the thesis comprises my original work, except where otherwise stated. All contributions from external sources have been acknowledged and explicitly referenced in the text. Some of the material discussed in the thesis has previously been published and presented at conferences. Appropriate references will be given in the relevant chapters.

Signed

Natalie Fecher

4th July 2014

1 Introduction

1.1 Making a case for facewear research

The present thesis is concerned with the effects of ‘facewear’ on speech. The expression ‘facewear’ is introduced in this context to refer to the various types of face-concealing garments and headgear that are commonly worn for occupational, recreational, religious and cultural purposes, and during the commission of crimes.

The research idea for investigating the impact of facewear on speech originally emerged from practical needs arising from casework conducted by forensic speech scientists. Against this background, the topic was for the first time addressed in 2008 by Llamas, Harrison, Donnelly, and Watt. These researchers opened up a then untouched field of study within forensic phonetics, and brought up a range of research questions which set the agenda for the current work.

To set the stage for further theoretical considerations and the empirical work introduced in later chapters, this chapter further exemplifies the use of the term ‘facewear’ in the context of the thesis. The motivations for conducting the research are then described. Specifically, it is acknowledged that the use of face and head coverings plays a role in forensic phonetic investigations at the present time, and that it is likely to be of relevance in future investigations. An outline of the thesis is also given at the end of this chapter.

1.1.1 Definition of ‘facewear’

The term ‘facewear’ will henceforth be used to refer to the large range of head and face coverings that people wear, more or less commonly, in everyday life. Face coverings fulfil very different purposes, are manufactured from different materials, and conceal the wearer’s head and/or face to varying degrees (half-face, full-face, mouth-only, etc.). The aim in the following sections is to illustrate the wide variety of masks that are regularly encountered in real-life communication situations.

Keeping in mind the forensic focus of the present work, the types of facewear that are typically worn during the commission of crimes (e.g. armed robberies and assaults) or in situations of public disorder (such as tumultuous demonstrations or riots), are of major concern here. A dip into the print and digital media suffices to get an idea of the large variety of face masks with which people choose to disguise their visual appearance on such occasions. They range from balaclavas, hooded sweatshirts, motorcycle helmets, scarves and bandanas (kerchiefs) wrapped around the neck and face, to thematic plastic or rubber masks (e.g. the Guy Fawkes mask, which is popular among anti-establishment protestors) and other forms of veiling (e.g. the white hoods which are utilised by members of the ETA group or the Ku Klux Klan). Of immediate forensic relevance are also the masks and materials that are involuntarily imposed upon the wearer, such as strips of duct tape forcibly adhered to a hostage's mouth during a kidnapping.[1]

However, this thesis takes a broader view of the subject. Here, 'facewear' denotes any type of garment or headgear that partly or fully conceals a person's head and/or face, and which is only *potentially* relevant in a forensic investigation. The latter implies that *any* spoken communication between individuals has the potential to lead to a situation with a legal aspect to it. For example, the ongoing political controversies about whether to prohibit the wearing of face-concealing clothing by Muslims in public places illustrate vividly how complaints can be based on claims about impaired speech communication on the part of a wearer of the face covering (for details on the '*burqa* debates' see §1.1.2.2).

According to this broader definition of the term, all and any face and head coverings worn for occupational purposes are of interest in this thesis. These comprise in particular the 'personal protective equipment' commonly used by police and law enforcement officers, firefighters, painters, miners, or workers on construction sites, in forestry, landscaping, and manufacturing plants. The professionals wear the masks in order to protect themselves from hazards likely to cause damage to their ears (e.g.

[1] We should also not forget the sometimes amusing, shocking, bizarre, genuinely creative or even artistic attempts of people to visually disguise themselves. Foraging in the media for the types of visual concealments chosen by bank robbers, shoplifters, and the like, one finds that practically no limits are imposed on the possibilities and the wearer's imagination.

noise above ~85dB SPL), face/skin (e.g. chemicals), or respiratory system (e.g. harmful gases or vapours). Examples include safety helmets with face visors, dust and surgical masks, oxygen masks and respirators, smoke hoods, welding masks, and a large range of hearing protection devices. A great variety of face masks can be found among medical doctors, nurses, and healthcare workers, as well as among military personnel. Examples of the former are surgical masks, disposable or reusable respirators, and oxygen or gas masks; examples of the latter are helmets, camouflage balaclavas, spandoflage head nets, strike steel (wire mesh) masks worn e.g. by soldiers, or breathing apparatus, gas masks and respirators worn e.g. by fighter aircraft pilots.

In addition, the face coverings of relevance in the present context include those worn for recreational purposes. Two health-related examples are the anti-pollution masks used by cyclists in cities with high levels of air pollution, and the surgical masks worn by members of the general public to reduce the risk of catching/spreading airborne diseases in densely populated areas in some East Asian countries. Protective face masks of all shapes and sizes are also found in sports, such as skiing, cycling, boxing, climbing, motorsports, fencing, or baseball, where hobby athletes and professional sportspeople alike wear a wide choice of helmets, face shields, balaclavas, and other specialised headgear. Moreover, there are no limits to the styles of plastic or rubber masks and wigs found at costume parties, or for entertainment on public grounds during festive seasons in many countries (carnival, folk festivals, etc.). Lastly, one should not forget the large variety of hats, caps, scarves, hooded sweatshirts, etc., which are worn for reasons of warmth, comfort, and fashion.[2]

[2] The present study does not consider facial make-up, even though the term ‘facewear’ might imply as much. It is not at all implausible to suppose that there might be a relationship between the amount of make-up a person wears and the sociolinguistic and sociophonetic variation exhibited in that person’s speech. In fact, this very thing has been shown to be the case e.g. by Mendoza-Denton (1997), who observed a correlation between the length of the eyeliner used by Latina gang members in California and the use of certain phonetic variants in the girls’ speech. In a similar vein, other researchers have looked at the relationship between the properties of a person’s speech and expressions of visual appearance (e.g. clothing, hairstyle), the purpose of which is to assist in the construction and projection of social identity and group membership (e.g. Eckert, 1996; Hay & Drager, 2007; Drager, 2009; Drager *et al.*, 2012). In this context, Hay & Drager (2007: 94) point out that ‘evidence of covarying linguistic and nonlinguistic factors makes it necessary to break down boundaries between studies of language, gesture, clothing, and other forms of social symbolism’.

Finally, this thesis takes into consideration the various kinds of veils, scarves, and headwear which some people choose to wear because of their religious and cultural significance. The designs, names, and purposes of use vary widely across societies and religions. Examples include the coverings worn by some Muslim women of Asian, Afghan, or Middle Eastern origin, like the *niqāb* (full-face veil that leaves a slit for the wearer's eyes), *burqa* (full-body cloth which is covering the entire head and body, with an optional fabric mesh obscuring the eyes), *ḥijāb*, *khimar*, *al-amira*, *shayla*, or *chador* (face, head, and chest covers of different styles). Other examples are the *tichel* (headscarf worn by some Jewish women), *keffiyeh* (headdress of some Middle Eastern men), *ghoonghat* or *dupatta* (long scarf worn over head and cleavage by some South Asian women), or turbans and turban-style headdresses.

The interested reader is referred to Winet (2012) for a survey of the historical practices of masking and veiling in Muslim and non-Muslim societies. Winet notes in this article that historically, 'veils' and 'masks' were similar in form, but over the centuries had begun to acquire opposing symbolic associations (at least in Western societies). While veils signalled 'high economic status, social respectability, and pious modesty' (Winet, 2012: 228), masks became associated with 'disguise, duplicity, sexual licence, and crime' (Winet, 2012: 228). There is no room to discuss the reasons for this divergence in the present work. But in view of the heated debates in the contemporary mainstream media about the 'ban of the *burqa*' from public places, it seems worth mentioning that over the course of this thesis, the term 'mask' will be used synonymously with 'facewear'. The same applies to expressions such as 'facial concealment', 'facial occlusion', 'facial obstruction', and the like. For stylistic reasons, all terms refer to *all* types of face coverings listed earlier in this chapter, including face-concealing veils such as the *niqāb*. It is recognised that the grouping of the *niqāb* along with the other face masks chosen for this study may cause offence to some readers. It should therefore be made clear that the author does not wish to imply any value judgements through the use of the employed terminology; all expressions are used without deliberately ascribing any moral, ethical or social values to them. On a final note, the reader is also referred to the definitions of 'mask' and 'veil' in the Oxford English Dictionary (2013), which exemplify the similarities and differences in the general use of both terms.

1.1.2 Motivation

1.1.2.1 Forensic phonetic casework involving facewear

One of the earliest known forensic phonetic cases dates from 30th January 1649, when King Charles I of England, Scotland, and Ireland was executed after having been found guilty for high treason. In spite of numerous theories that have been put forward since, till this day neither the identity of the executioner who beheaded him, nor that of his assistant, are known for certain. This arises from the fact that, as was common at the time, the faces of both men were hidden by a face mask to protect them from reprisal for killing the king. One of the suspects charged in the investigation following the decapitation 11 years later was Sergeant William Hulet. The allegation was largely based on testimony given by Officer Richard Gittens, who claimed to have recognised the voice of the executioner when he (the executioner) spoke to the king on the scaffold (Hollien, 1990; Solan & Tiersma, 2005; Foulkes & French, 2012). The relevant sections of Gittens' statement (extracted from Howell *et al.*, 1810: 1186f.) read as follows (all spellings as per original):

Gittens.	[…] Hulet (as far as I can guess) when the king came on the scaffold for his execution, and said, Executioner, is the block fast? then he fell upon his knees.
Counsel.	Who did?
Gittens.	Hulet, to ask him forgiveness; by his speech I thought it was he; […]
Counsel.	Did you know his voice?
Gittens.	Yes, sir.
Counsel.	Did you mark the proportion of his body, or his habit, what disguise he was in?
Gittens.	He had a pair of freeze trunk breeches, and a vizor, with a grey beard; […]
Hulet.	I desire as to this witness; he doth alledge that he and I were serjeants in one company, which I deny; he was not in that company I was in; I desire to know of him how he comes to know that I was there at that time.
Gittens.	By your voice.

The regicide of King Charles I, the subsequent storm of inquiries about who was ultimately responsible for it, and the trial of William Hulet, can be considered as one of the earliest documented forensic cases involving speech produced through facewear. The incident portrays a case where a witness claims to have identified the voice of a familiar person whose face was disguised by some form of masking.

To this day, forensic speech scientists are regularly faced with related issues concerning the recognition of an individual by voice alone. The field of forensic speech science is introduced in more detail in §2.2.1. For the time being, it suffices to acknowledge that the current research aims to meet a practical need. This is to do with providing experimental data on which forensic practitioners can ground estimates of the influence facial disguise may have on the reliability of (lay and expert) evidence. If the speech in dispute in an investigation was produced through facewear, the expert should be prepared to take that knowledge into account as a potential influencing factor during his/her analysis of the evidential speech material. The contribution of facewear research within forensic phonetics is further highlighted in §2.2.2 and §7.2.

The relevance of the present research to casework carried out by forensic speech experts was the primary motivation for initiating the study presented in this thesis. Indeed, forensic phonetic casework that involves some form of facial concealment on the part of the talker (or listener, for that matter) is not exceptional, but is quite a frequent occurrence. This is affirmed by Peter French (chairman of *J P French Associates*, York) based on his experience with working on thousands of forensic phonetic cases from around the world (Peter French, personal communication, 2nd May 2013). It is also accentuated by the fact that in the course of working on this thesis, the author has on three occasions been consulted directly as an expert on the subject. The requests were made by forensic practitioners working in established forensic laboratories in England and Germany. One was concerned with a case in which the point at issue was whether the speaker's mouth could have been taped closed by a piece of adhesive tape at the time a call to the emergency services was made. Another one involved the question of whether the visual identification of a face which is partly concealed by a balaclava is considered feasible. The third task assigned to the author involved the design and execution of an acoustic

reconstruction experiment, which tested the influence of speech produced through a motorcycle helmet on the reliability of earwitness testimony (specifically, whether it would have been possible for the witness to have identified the speaker by his voice, as well as having been able to identify the words being spoken).

1.1.2.2 Contemporary debates on facewear use in public

Llamas *et al.* (2008) state that their original motivation for conducting facewear research stems from the public debates about whether to ban face-concealing Muslim clothing from public places (streets, parks, civil institutions, etc.). These were ongoing at the time across wide parts of Western Europe, and continue till this day. To recall, the religious and cultural dress code under dispute consists of lightweight garments worn over a woman's head and face, with some styles (*niqāb*) leaving a small slit for the eyes, and others covering the wearer's eyes with a semi-transparent fabric mesh (*burqa*).

The extensive online and offline media coverage, and the numerous internet forums, blogs, TV and radio broadcasts devoted to the topic demonstrate that the proposals of various state governments for legally prohibiting the use of such clothing in public has been (and still is) the cause of much heated discussion. The political and socio-cultural controversies often centre on religious and personal freedom, female equality, a presumed 'clash of civilisations' between the 'secular West' and 'religious East', related questions about multiculturalism, integration, and minority rights, as well as safety concerns and fear of terrorism (Winet, 2012; *BBC News*, 2013a). The motives cited for enacting a ban of headscarves and face veils from certain public domains vary widely across nations, but an argument frequently put forward is that the inability to (visually) identify a person by his/her face poses a security risk (Winet, 2012). This rationale is fuelled by reports of perpetrators committing crimes while wearing a *burqa* to evade recognition (*BBC News*, 2013b, 2014).

As of today, some countries have already passed laws that officially forbid the veiling of the face in public, like France (in 2010) and Belgium (in 2011), while others are still debating similar nation-wide legislation (e.g. Italy) and/or have more limited prohibitions, such as Australia, Canada, Egypt, Germany, Kosovo, Netherlands, Norway, Singapore, Spain, Sweden, Syria, Tunisia, and Turkey (Winet, 2012; *The World Post*, 2013). Winet (2012: 239) points out that some laws and regulations specifically address the face veil worn by Muslim women, and others regulate the use of religious attire or face-concealing garb and headgear more generally; some encompass all public spaces, and others are restricted to certain environments (especially legal and educational institutions); some are imposed by national governments, and others by municipalities or individual institutes. The most prominent example of the latter are school uniform policies, which empower school supervisors and education boards to expel students or fire teachers who infringe the provisions. For example, both the British and German governments have thus far not passed national laws restricting religious dress in public. However, following several high-profile court cases, Britain has ruled that dress codes should be at the school's discretion (*BBC News*, 2007a, 2007b, 2013c), and Germany transferred the right to impose restrictions on clothes worn by school personnel to the 16 state governments (see 'Kopftuchstreit', e.g. in *Süddeutsche Zeitung*, 2010).

But requests for regulating by law the use of facewear in public are by no means confined to the wearing of face-concealing clothing in accordance with religious beliefs. Several countries have long since taken legal measures against the use of any face masks worn for reasons of anonymity. As Winet (2012) notes, the existing laws and regulations can in principle be applied by law enforcement officials to all (religious and non-religious) garments that obscure an individual's face. The need for legally controlling facewear use for anonymity purposes arises in situations that require the (visual) identification of the wearer, for example, demonstrations and protests leading to incidents of material damage or offences against the person. In their strongest form, so-called 'anti-mask laws' prohibit the wearing of any face covering in public. This primarily aims to criminalise actions of masked individuals in the above situations (Simoni, 1992; Harvard Law Review, 2004; Winet, 2012). The interested reader is referred to Appendix A for several excerpts from anti-mask legislation in different countries (e.g. the US, UK, and Germany). The reasoning

behind the decision made in different nations of whether or not to outlaw the wearing of face coverings in public, and the arguments raised both in favour and at the expense of anti-mask laws and Muslim face veil bans, cannot be laid out in further detail in the context of this thesis (for further discussion see e.g. Winet, 2012). However, in view of the scope of this thesis, namely to examine facewear effects on speech, one argument that is frequently delivered in support of the headscarf ban is of particular interest, and therefore deserves some attention.

Opponents of the Muslim face veil often claim that obscuring one's face complicates 'face-to-face' communication with another person. As Llamas *et al.* (2008) point out, the argument that the veil degrades the *acoustic/auditory* speech signal – i.e., that it causes difficulty in hearing and understanding the veiled person – is less often put forward than the argument that the veil hinders the extraction of *visual* information from the talker's face. Specifically, '*burqa* ban' supporters frequently argue that this type of face covering obscures facial speech cues that are important for processing conversational speech and for recovering the talker's intended message, that it prevents the interlocutor from reading facial expressions and acknowledging emotional reactions, and that it prohibits socially- and conversationally-relevant eye contact between interlocutors (at least in case of veils that also cover the eyes).

This line of argumentation is most commonly found in the context of legal and classroom discourse. There have been several cases where a female witness giving evidence in court was asked to remove her face veil under the pretext that the judge, magistrate, or other judicial office-holder 'could not hear her properly' (*BBC News*, 2006), that the veil did not make it possible for the judge, jury, and lawyers to 'see and assess her responses' (Casciani, 2013), and that it hindered 'openness and communication' (Casciani, 2013). In the UK, decisions of this kind are currently directed by the *Equal Treatment Bench Book* circulated by the Equal Treatment Advisory Board Committee of the Judicial Studies Board (see UK Judicial Studies Board, 2013). These guidelines (revised November 2013) acknowledge the difficulty and sensitivity of regulating religious dress code in court. They advise that it is for the judge to decide whether any steps are necessary to ensure effective participation and a fair hearing for the woman wearing the *niqāb* and for all other participants in the proceedings. However, it is generally advised that witnesses who choose to cover

up should not be requested to remove the clothing in court, for reasons of preserving the individual's right to freedom of religious practice (see also Kirk, 2013).

Furthermore, there have been many instances where young women were requested by school authorities to unveil in classrooms or in order to prove their identities when sitting for state examinations, or where teachers who did not comply with the schools' dress codes were ordered to leave their placement schools (Todd, 2010). The reasons related to interpersonal communication that are given in this context are generally similar to the ones mentioned above. For example, it has been argued in the past that 'the full face veil hampered communication' between teachers and pupils (*BBC News*, 2007a), that veils 'make communication and learning difficult' (*BBC News*, 2007b), that face coverings which 'prevent teachers from seeing pupils' facial expressions are "not suitable in school"' (*BBC News*, 2013c), and that being able to extract students' facial reactions is a 'key element in effective classroom interaction' (*BBC News*, 2007b).

1.1.2.3 Proliferation of audio-visual surveillance

Aside from the issues raised in the previous section, the present research on speech produced while the talker is wearing a face covering was stimulated by the supposition that the number of forensic cases involving the analysis of speech produced through facewear is likely to be significant, or even to increase, in the future. The argument put forward in this regard is that due to reinforced privacy concerns associated with the recent rise of visual surveillance, there may be a greater necessity, or preference, for individuals to disguise their visual appearance during certain (potentially illicit) activities. This hypothesis, albeit speculative, is backed up by recent developments in the public and private sector. Some of these developments are described in the following sections.

Today, visual surveillance surrounds us more or less constantly. This trend is most vividly demonstrated by the rapid increase in the number of CCTV ('closed circuit

television') cameras installed since the early 1980s. The highest number of cameras per citizen can be found in the United Kingdom. Here, nearly every step citizens take in public spaces (streets and pavements, retail and commercial premises, shopping malls, public transport, train and bus stations, universities, hospitals, airports, etc.) is caught on camera, for reasons of monitoring, surveillance, safety and security. The often-cited figure of around 4.2 million operational cameras in the UK alone (published by the Home Office in 2003) was recently claimed to be an exaggeration. Statistics provided by the Association of Chief Police Officers state there to be 'only' around 1.85 million CCTV cameras in the UK (Gerrard & Thompson, 2011). Whatever the number, it is probably safe to concur with Ball *et al.* (2012: 2), who say that there has been a 'momentous expansion and intensification of surveillance in almost all institutional spheres of contemporary existence'.

The results of research into the effectiveness of CCTV systems as a crime prevention measure are controversial. However, by and large studies report a reduction in the number of crimes in experimental areas where cameras have been installed (see e.g. Armitage, 2002; Gill & Spriggs, 2005; Welsh & Farrington, 2007). Gill & Spriggs (2005) further remark that the awareness of cameras among the general public has increased over the years. Armitage (2002) claims CCTV to be a useful tool for the deterrence of (potential) criminals, who may more thoroughly assess the risks of offending in a location where CCTV is in operation. In the context of this thesis it can be argued that offenders may 'compensate' for the increased risk of being caught by disguising their visual appearance so as to not be recognisable from the footage. Self-evidently, one way of doing so is by means of wearing face-concealing garments and headgear.

In fact, the above argument holds for any individual who is aiming to secure privacy by wearing a face covering, as for example, when participating in nonviolent demonstrations or street protests. Furthermore, the problem of being recognisable from video footage is not confined to CCTV images. It also applies to recordings on live broadcast television, and to videos and photographs taken with personal electronic devices, such as smartphones, mobile phones, and digital cameras. As with CCTV, the number of the latter devices has increased tremendously in recent years. *Ofcom*'s 10th annual Communications Market Report states that in 2013, 51% of

adults owned a smartphone in the UK (compared to 27% in 2011), and that 94% of adults possessed a mobile phone. The majority of smartphone users (especially teenagers and young adults) reported to having their device constantly switched on, sometimes even in places where they are asked to turn it off.

The images captured with aforementioned devices are nowadays often shared via social media (e.g. *YouTube*), and the devices have been identified as being actively involved in crowdsourced law enforcement. This arises from the fact that short clips of public disturbances, shop lootings, fatal attacks, violations of human rights, etc., can be of great value in a forensic investigation (Firth, 2001; Hogan, 2003; Graham-Rowe, 2006a, 2006b). This trend has been succinctly illustrated by the *New Scientist* article entitled 'Smartphone surveillance: The cop in your pocket', which states that 'we are all set to gain unprecedented crime-fighting abilities', and that '[w]hile many of us use smartphones to keep our social lives in order, they are also turning out to be valuable tools for gathering otherwise hard-to-get data' (Fleming, 2011: 1).

While smartphones and other technologies mediate our lives already, it is predicted that cameras will be integrated in our society even more in the future (Mann *et al.*, 2003; Mann, 2013). The exponential advances e.g. in computer vision and information technology over the last decades already enable us to purchase 'wearable' technologies, i.e., body-borne computers and miniature electronic devices, which often incorporate video function (e.g. *Google Glass*). The advantages that the steady technological progress offers in terms of crime prevention and investigation are now indisputable. But it is no surprise that this is at odds with the prevailing privacy concerns related to (audio-)visual surveillance forced upon private individuals 'from above' (by means of ever smarter surveillance systems) and 'from below' (by means of privately-owned portable recording devices).[3]

Following these introductory considerations, and the definition of the term 'facewear', the next section provides the reader with an outline of the thesis.

[3] Surveillance 'from below' has been termed *sousveillance* by Steve Mann (see e.g. Mann *et al.*, 2003), which is playing on the similarity but opposing meanings of the prefixes *sur* (from French, 'over'/'above') and *sous* (from French, 'under(neath)'/'below').

1.2 Thesis outline

Chapter 1 familiarised the reader with the general theme of the present work by firstly providing a definition of the term 'facewear'. The remainder of Chapter 1 presented the primary motivation for carrying out facewear research, namely its practical application in forensic casework conducted by speech scientists. A brief review of 'anti-mask' legislation and the contemporary debates on the Muslim face veil ban was then given, and the role of audio-visual surveillance (and sousveillance) as an integral part of modern society was portrayed. The latter aimed at contextualising the research, and reinforced the assertion made that the wearing of face and head coverings in public places and during public events is likely to play a role in future forensic phonetic investigations.

Chapter 2 introduces the research directions taken in this thesis. These relate to the production, acoustics, and perception of consonants that are produced while the talker's face is disguised by a mask. Furthermore, Chapter 2 lays out the approach chosen to investigate facewear effects on consonants, along with some of the difficulties encountered while doing so. The second part of Chapter 2 gives a concise overview of the field of forensic speech science, and lists the various factors known to influence speaker recognition performance by expert and lay witnesses. In this context, the contribution of facewear research within forensic phonetics is outlined. Lastly, Chapter 2 provides a theoretical account of preceding research on the influence of facewear on speech. Despite its immediate forensic relevance, there exists surprisingly little research on the topic from a forensic phonetic perspective. As previously mentioned, the pioneering work on the acoustic and perceptual effects of forensically-relevant face coverings on speech was carried out by Llamas *et al.* (2008). Their study set the agenda for the empirical work presented in the thesis, and is therefore discussed more thoroughly. Finally, other forensically-motivated research on facewear effects on speech is surveyed, and a range of thematically-related studies, which examined the impact of different types of face masks on speaking and listening more generally, are summarised.

To expand on this underexamined field of study within forensic speech science, several experiments were set up to explore the manifold effects that facewear can

have on the production, acoustics, and auditory(-visual) perception of consonants. To be in a position to conduct this research, it was necessary to collect appropriate speech material that could be used for experimentation. This was vital because no speech database already existed which a) provided sufficient control over the speech material, and the acoustic environment the material was elicited in, b) was comprised of audio and video recordings of speech produced by talkers whose faces were concealed by facewear at the time the speech was produced, and c) included a comparatively large variety of face and head coverings. **Chapter 3** describes the design of a speech corpus consisting of high-quality audio and video recordings that fulfil the above criteria. The data were collected in recognition of the fact that the occlusion of a talker's face while s/he is talking is likely to have a combined articulatory, acoustic (+ auditory), and where applicable, visual impact on the production and perception of the speech signal. The corpus is henceforth referred to as the 'Audio-Visual Face Cover' (AVFC) corpus.

Chapter 4 presents two experiments which focus on the acoustic-phonetic analysis of consonants produced through facewear. The first experiment (Experiment 1) deals with the analysis of intensity and spectral measures of the voiceless fricatives /s/, /ʃ/, /f/, and /θ/. The second experiment (Experiment 2) attends to intensity, spectral and temporal properties of the voiceless plosives /p/, /t/, and /k/. The questions addressed in the study are: does facewear change selected intensity, temporal, and spectral measures of fricatives and plosives when the consonants have been produced while the talker's face is disguised by facewear? If so, are the two classes of fricatives (sibilants and non-sibilants) differently affected by facewear? Correspondingly, to what degree and in what manner does facewear alter the acoustic characteristics of plosives? And which type of face covering has, by and large, the most deleterious effect on the acoustics of the speech sounds? Chapter 4 describes the motivations for analysing fricatives and plosives, along with their most relevant articulatory and acoustic properties. This is followed by the presentation of the applied methodology, the statistical analysis of the data, and the discussion of the most important findings.

Chapter 5 addresses a set of research questions related to the auditory(-visual) perception of consonants produced through facewear, such as: does facewear hinder the identification of consonants when these have been produced through facewear?

Do lay listeners more accurately identify the consonants when they can watch the talker's articulating face and hear the talker's voice (compared to only hear the talker's voice)? Can listeners extract visual speech information from the talker's face even when the face is partly or fully concealed? To approach these questions, a consonant identification study comprised of two experiments was carried out. The first experiment (Experiment 3) addresses consonant identification in quiet listening conditions (studio-quality recordings), while the second experiment (Experiment 4) is concerned with testing the same set of stimuli when the original soundtracks were intermixed with babble noise. To place this work in a broader theoretical context, Chapter 5 opens with an introduction to the research area of auditory-visual speech processing. After that, the methodology employed in Experiments 3 and 4 is described, and a discussion of the experimental results is provided. Subsequently, a phonetic feature analysis (using the signal detection measure *d-prime*) is presented, which offers an examination of the types of perceptual errors that participants made. Chapter 5 closes by discussing all findings in relation to the literature.

Chapter 6 deals with the perception of the indexical (talker-specific) properties of consonants. Specifically, Experiment 5 examines the ability of phonetically-untrained listeners to determine whether short samples of speech (CV syllables) have been spoken by the same talker or by two different talkers. This is tested under the condition that the speech material has been produced when the talkers' faces were undisguised, when the talkers were wearing a motorcycle helmet, and when they were speaking while their mouths were taped closed. Experiment 5 seeks to provide answers to the questions: can lay listeners correctly determine whether two samples of speech originate from the same talker or from different talkers when all the listeners have available for comparison are CV syllables? Does facewear change the talker-specific properties of speech? Does facewear negatively impact on talker discriminability? Furthermore, the study builds on and extends the findings from previous research (introduced at the outset of Chapter 6), which has shown that the processing of indexical properties of speech can be significantly affected by the linguistic content of the speech signal. Against this background, the study explores whether the segmental content of the test samples (here, the six consonants /t p s f n m/ embedded in the CV syllables) had an effect on the listeners' performance in distinguishing between unfamiliar talkers. That is, do some consonants possess

greater talker-discriminating potential than others? Chapter 6 offers a report of the applied methodology and experimental results, and closes with a discussion of the main findings in view of the literature.

The concluding chapter of this thesis, **Chapter 7**, summarises the content of the preceding chapters, and in particular spells out the core findings from the empirical studies presented in Chapters 4 to 6. Thereafter, Chapter 7 highlights the practical implications of the research in the context of casework carried out by forensic speech scientists. It is emphasised that facewear effects should be taken into account by practitioners when interpreting the results of their acoustic and aural analyses of evidential speech recordings, and when evaluating the reliability of lay earwitness statements. Finally, Chapter 7 provides some ideas and directions for future research, which is believed to be beneficial in strengthening our current understanding of the effects of facewear on speech acoustics and perception.

2 Theory and literature review

2.1 Research directions

The present chapter offers a theoretical foundation for research on the effects of facewear on speech. To begin with, the chapter introduces the theoretically-feasible research directions more broadly. After this, the focus is narrowed down to the areas of research that are addressed in this thesis. These are to do with the production and acoustics of speech on the part of the mask wearer, and the perception of the mask wearer's speech on the part of the (unmasked) listener. Furthermore, the approach that was chosen to investigate the effects of facewear on speech, along with some of the challenges and difficulties that were encountered while doing so, are outlined.

2.1.1 Overview of 'facewear speech'

The fundamental question addressed in this thesis is: does facewear influence the way that speech is produced, transmitted, and perceived? The research presented in the upcoming chapters is intended to offer the first large-scale study of the (likely) effects of different types of facewear on speech.

In this context, facewear can be pictured as a 'physical barrier' that is placed somewhere along the 'speech chain' between two interlocutors. To help to understand how facewear affects the speech communication process, it is important to assess its effects on both the talker and the listener. Speech that is produced while the talker is wearing some kind of facewear is henceforth referred to as 'facewear speech'; speech produced by the same talker when s/he is not wearing facewear is referred to as 'control speech'.

There are various angles from which the topic could be approached. Before moving on to introduce the research directions taken, the different possible perspectives are outlined by way of a general overview. Owing to time and space constraints, only a fraction of the viable research questions can be addressed in this thesis.

For illustrative purposes, let us assume that ‘Interlocutor A’ represents a person who is wearing some sort of face covering, while ‘Interlocutor B’ represents a person whose face is *not* concealed in any way. In this communication scenario, the following questions concerning Interlocutor A can be asked:

A1. Is the production of A’s facewear speech different from that of A’s control speech?*
A2. Are the acoustic properties of A’s facewear speech different from those of A’s control speech?*
A3. Is A’s perception of B’s speech different when A is wearing facewear?
A4. Is A’s perception of his/her own voice different when A is wearing facewear?

With respect to Interlocutor B, the following can be considered:

B1. Is B’s speech produced in response to A’s facewear speech different from B’s speech produced in response to A’s control speech?
B2. Are the acoustic properties of B’s speech produced in response to A’s facewear speech different from those of B’s speech produced in response to A’s control speech?
B3. Is B’s perception of A’s facewear speech different from B’s perception of A’s control speech?*
B4. Is B’s perception of his/her own voice different when B is responding to A’s facewear speech (as opposed to responding to A’s control speech)?

The research directions marked with an asterisk (*) are dealt with further in this thesis. They are introduced in more detail in §2.1.2.

Questions A3 and A4 (perception of the interlocutor’s speech and of one’s own voice when wearing facewear) mainly relate to the types of masks which also (or exclusively) cover the ears. Most will have the experience that concealing one’s ears may cause difficulties in hearing another person, and also, that the auditory feedback of one’s own voice is altered to some degree (e.g. it may sound louder or somewhat

dull) when the ears are covered up. Examples of common 'earwear' are helmets used to ride a motorbike, woollen or fleece hats to protect the wearer from the cold, and also the various audio playback and hearing protection devices placed on top of or inside the ear canal (e.g. hands-free telephone headsets, in-ear headphones, noise-cancelling earplugs).[4] Previous research on the effects of earwear on speech perception (and production) will be introduced in §2.3.3, and opportunities for future research in this domain will be proposed in §7.3.

The research questions B1 and B2 refer to a (non-disguised) talker's verbal behaviour in response to hearing (and seeing) an interlocutor who is wearing a face covering of some sort (Interlocutor A in the above example). Here, 'verbal behaviour' denotes the changes to the talker's own speech productions and the resulting speech acoustics when s/he is communicating with a (disguised) interlocutor. These changes may become apparent, for example, via an increase in speaking volume, a reduced speaking tempo, a more exaggerated way of articulating, or in adjustments to the talker's interpersonal communication strategies (e.g. turn-taking signals). The latter applies in particular to the types of masks that hinder eye contact with the interlocutor and the extraction of facial expressions. Note that research question B4 is closely related to questions B1 and B2, in that some degree of monitoring of one's own voice/speech is always necessary when aiming to produce intelligible speech. This research direction also refers to the assumption that some talkers may intentionally adapt their verbal behaviour, while others may not be conscious of the fact that they modify their usual way of speaking.

The adaptations of a talker's verbal behaviour to a masked interlocutor may be triggered by a range of emotional and attitudinal reactions, or certain expectations and biases, towards the person wearing a particular face covering. As was discussed in connection with the '*burqa* debates', some types of facewear may lead to assumptions of reduced intelligibility of the speech produced by the wearer. This likelihood was affirmed by a short questionnaire administered as part of the research

[4] On a side note, perceiving the interlocutor's speech or one's own voice differently when wearing earwear may in turn change one's own speech productions and acoustics. Hence, the questions A3 and A4 ought to be studied in conjunction with A1 and A2.

to be presented in this thesis. It was found that participants in a listening test (in which they were exposed to facewear speech and had to make certain judgements about it) assessed the intelligibility of facewear speech as lower *before* taking the test than after having completed the test. Further details about the questionnaire and suggestion for future research in this area will be given in §7.3.

2.1.2 Focus of the thesis

The research directions taken in this thesis are concerned, firstly, with the way speech is produced when the person talking is wearing facewear. Secondly, it investigates the acoustic properties of the mask wearer's speech, and thirdly, it deals with the auditory(-visual) perception of facewear speech. In the following sections, these three research directions will be introduced in more detail.

2.1.2.1 Speech production

The way we produce speech is likely to be altered when a mask is covering our face. This claim seems plausible from our personal experience and expectations (for example, when imagining a scarf tightly wrapped around our neck and lower half of the face, or the solid shell of a tight-fitting motorcycle helmet limiting our head and face movements). In such situations, the 'default' motor activity of certain active articulators (such as the lips), normal facial surface behaviour, and/or natural jaw motion, may be impeded to some degree. For instance, when a mask applies (some) pressure on the outer surface of the face, muscle contractions in and around the lips may be interfered with. Consequently, the relative positions of the upper and lower lip may be changed. This in turn may impair the forming of a given talker's typical bilabial closure, which is necessary for the production of consonants like /p/ and /m/.

Facewear-induced modifications to speech articulation of this kind can be considered to be passive. This means that they do not comprise any voluntary involvement of the talker as such, but occur sporadically as a consequence of the mask getting in the way of the normal functioning of the articulators. In addition, a talker might actively compensate for wearing facewear. This may occur in response to merely the anticipation of being less well understood, or to compensate for the lack of facial speech cues. Hence, facewear speech could perhaps be characterised by a more pronounced or even exaggerated manner of speaking, or by increased vocal effort (involving pitch, loudness, and duration; see e.g. Traunmüller & Eriksson, 2000; Jessen *et al.*, 2005).

In the present thesis, the modifications to speech production triggered by the wearing of facewear are not examined as such, but inferred from the results of acoustic-phonetic measurements of facewear speech, and from auditory judgements of the same (i.e., careful listening and observing). Knowledge derived from general phonetic theory, including speech perturbation and compensation studies (see e.g. Gracco & Löfqvist, 1994; McFarland & Baum, 1995; Baum *et al.*, 1996; Ito *et al.*, 2000; Brunner, 2009; Ménard *et al.*, 2013), will be of particular value in this respect. It is anticipated that the findings will support our understanding of the facewear-activated changes to the acoustic speech signal and to the perception of speech that is produced while the talker's face is concealed.

2.1.2.2 Speech acoustics

Building on the considerations discussed in the preceding section, the question arises of whether the acoustic properties of facewear speech differ from the acoustic characteristics of control speech produced by the same talker. Based on research by Llamas *et al.* (2008), which is presented in more detail in §2.3.1, it is hypothesised that the modifications to the acoustic signal brought about by facewear will originate principally from two sources.

Firstly, it is common knowledge in phonetic theory – e.g., the source-filter theory of speech production (Fant, 1960), or Steven's quantal theory of speech (Stevens, 1972; Stevens, 1989; Stevens & Keyser, 2010) – that even minor modifications to the articulatory gestures during speech production may alter the resultant acoustic signal. Hence, even slight repositioning of the talker's articulators while s/he is speaking through facewear is likely to give rise to prominent changes to the acoustic properties of the produced sounds. For example, the mechanical perturbation (impeded lowering) of the jaw provoked by, say, a motorcycle helmet, may result in a reduction of the speaker-specific range of the first formant of open vowels (see e.g. Clark *et al.*, 2007).

Secondly, acoustic facewear effects will arise simply by virtue of a physical obstruction occluding the talker's face. When a fabric or other material is covering the mouth and nose, the propagation of the sound wave will be hindered, and the sound energy of certain spectral components of the signal will be 'lost' (absorbed). Moreover, when the air molecules hit the obstacle outside the mouth, additional turbulences may be created. This may auditorily become apparent as 'hissing' or 'whistling' sounds. The degree of such interference will be determined by the sound-absorbing characteristics of the particular facewear material, and by the fit of the mask around the talker's head/face. For a simple demonstration of the acoustic absorption effect, one just ought to imagine a talker holding a hand closely in front of his/her mouth while speaking. Most readers will know from experience that this will cause the talker's voice to sound slightly 'muted', 'muffled', or 'dull'.

The acoustic facewear effects will be addressed in this thesis by taking both sources of acoustic change into consideration. Chapter 4 presents an acoustic-phonetic analysis of selected speech sounds which were produced through various forensically-relevant face coverings. The comparison of intensity, spectral, and temporal measures taken from facewear speech with the same measures made from control speech aims to provide valuable insights into the acoustic modifications to the speech signal that can (and in practice should) be expected when a talker's face is concealed by facewear.

2.1.2.3 Speech perception

The third research direction concerns the question of whether phonetically-untrained listeners can actually hear the differences between facewear speech and control speech. At present, two answers to this question seem plausible. On the one hand, the (possible) articulatory and acoustic changes to speech caused by facewear might be only minor ones that have no perceptual consequences for listeners (who simply 'ignore' them or factor them out). On the other hand, the speech signal could potentially be modified to the extent that speech processing is impaired on the part of the listener.

Over the course of this thesis, two speech perception experiments testing for both alternatives will be discussed. The first study examines the identification of consonants produced through facewear. The stimuli used here are presented in quiet and noisy listening conditions and under the condition that the participants could either only see, or see and hear, the talker (see Chapter 5). The second study tests lay listeners' ability to distinguish between the voices of two talkers who are either wearing or not wearing facewear (see Chapter 6). The goal of both studies is to evaluate whether the perceiver's performance in these tests changes – for good or bad, or not at all – when a mask interferes with the processing of the talker's speech.

In this context, the multimodal nature of the present topic will be introduced. During natural face-to-face communication, a wide range of conversationally-relevant *visual* speech cues are available to interlocutors. By watching the talker's face, head, and hand movements, listeners extract not just the linguistic message, but also information about the talker's identity, emotional and physical state, and so forth. Simply speaking, our overall impression of a person, and our understanding of that person's spoken message, is determined both by what we hear and what we see. In the current work, the focus will be exclusively on visual information that can be extracted from the talker's face, and which informs the listener about the segmental content of the produced speech (consonants and vowels). The field of auditory-visual speech processing will be introduced in more detail in §5.1.1.

2.1.3 Research approach

To test experimentally for the occurrence of facewear effects on speech production, acoustics, and perception, numerous methods and procedures could be chosen, and many questions would still be unanswered. Bearing in mind that there exists virtually no previous forensic research on the topic (other than Llamas *et al.*, 2008), the present work ought to be considered a first step towards a better understanding of the influence of face coverings on speech. It is anticipated that the current research will establish a theoretical framework for future research, and provide some solid foundations concerning the effects that can and should be expected when facewear is involved in the speaking and listening process.

When planning the course of action for this work, many compromises were necessary. The intention (and associated difficulty) was to carry out research within the bounds of scientific possibility, as well as within the limits of admissibility of the resultant research findings among forensic and judicial practitioners, and in court. Regarding the former, the aim was to set up a range of experiments that would address narrowly-defined research questions, enable careful control over the experimental designs, and generate reproducible results. To meet these goals, the acoustic study follows established procedures borrowed from acoustic phonetics, and the perception studies adopt classic experimental designs employed by psychologists and psycholinguists in behavioural studies of language processing. The latter procedures enable the researcher to keep constant all (or at least many) dimensions of the object of investigation, and only manipulate the dimension(s) of interest. This has clear advantages in terms of controlling and interpreting the data. However, such procedures can be difficult to apply to multi-dimensional phenomena – such as the human voice – and may come at the expense of ecological validity.[5]

[5] Ecological validity is often associated with the generalisability of the findings from a research study to the 'real world'. Here, it refers to the question of whether we can extend the results emerging from speech production and perception experiments conducted in a research laboratory to the way people produce and process speech in natural communication environments. The ambition to perform research with high ecological validity is particularly pertinent to forensic speech scientists, because the conditions in which relevant speech material is produced and/or witnessed often deviate radically from the conversational environments that people encounter on a day-to-day basis.

In addition, the work aspired to meet the requirements imposed on all research conducted by forensic phoneticians, linguists, and acousticians, which are concerned with the admissibility of the generated research results among the relevant communities. The acceptance (and comprehension) of the scientific work carried out by the expert is generally higher on the part of the judicial audience when the research clearly demonstrates a 'real-world' application in terms of the research questions asked, the speech material examined, the subjects tested, and so forth. Typically, the research carried out by forensic speech scientists concerns the factors known (or expected) to influence the production, acoustics, and perception of speech in forensically-relevant situations. The overall goal is to produce research results that can serve as a reference in future casework carried out by the analyst him-/herself, or by fellow experts.[6] The difficulty that arises in this respect is that keeping the degree of forensic realism of the research as high as possible sometimes inevitably comes at the expense of experimental control.

The speech data incorporated in the current experiments derive from audio and video material which was recorded while the talker's face was actually disguised by facewear at the time the speech was produced (see Chapter 3). To that extent, the data reflect the talker's speech productions as they 'naturally' occur while s/he is talking through facewear. In other words, the 'real-life' aspect of the present work *was* that the speech material was elicited from talkers whose mouth or entire face were actually concealed while talking.

The approach applied could be described as 'bottom-up'. It was decided to start from a relatively low linguistic level, and to centre the examination of facewear effects on a basic (albeit not undisputed) unit of speech, the phoneme. Specifically, facewear speech is studied by observing facewear-induced acoustic and perceptual changes to spoken English consonants. Acoustic facewear effects are explored by measuring acoustic-phonetic properties of consonants. Perceptual facewear effects are examined by testing listeners' performance at identifying consonants produced through

[6] The specific research questions often emerge from cases that the analyst has previously worked on, and on occasion studies are carried out as an integral part of casework (e.g. acoustic reconstructions).

facewear, and at distinguishing between different voices based on short consonant-vowel sequences. Confining the analysis to the consonant level was considered worthwhile for the following three reasons.

Firstly, the study of individual segments seemed justified because forensic practitioners commonly reduce speech into its component units when they analyse evidential speech recordings (Gold & French, 2011; Foulkes & French, 2012; for further details about the analytical procedures that are regularly applied by forensic phoneticians see §2.2.1). The examination of a set of consonants seemed favourable because of their energy distributions across a wide range of frequencies (including ranges higher than those of the third formants of vowels, for example). Previous research (especially Llamas *et al.*, 2008) suggests that face masks can influence the acoustic speech signal particularly in these higher frequency bands. This makes consonants especially prone to acoustic modifications caused by facewear, and consequently, to misperception by listeners. Even when no facewear is involved, consonants are already known to be less robust (e.g. in noise) than are vowels or rhythmic features of speech (see e.g. Fraser, 2003).

Secondly, it seemed more beneficial to begin the investigation into facewear effects from a rather low linguistic level, and to systematically tease apart the effects of facewear on the production, acoustics, and perception of individual phonemes. By narrowing down the analysis to the consonant level, it was possible to extract some of the articulatory, acoustic and perceptual effects on speech caused *by facewear* – and not by other contingent factors (including e.g. lexical or syntactic predictability). If the research were to show that facewear has an effect at the level of the individual consonant, it could be concluded from this that human listeners are sensitive even to the fine-grained acoustic differences that facewear brings about. To ultimately understand how (if at all) facewear affects the lexico-semantic processing of spoken utterances, future research should focus on meaningful words/sentences and natural conversations involving facewear use, and will ideally simulate forensically-relevant communication scenarios (e.g. in the form of mock voice line-ups).

Thirdly, the experiments can be linked, to the extent that the same object of investigation (here, spoken English consonants) will be viewed from different angles.

That is to say, the study looks at the way that the consonants are produced, at their acoustic properties, at how well they can be identified by lay listeners, at how much talker-specific information they convey, and, most importantly, at the extent to which these properties change when the consonants are produced while the talker's face is disguised by facewear.

In summary, the reader has so far been presented with the research directions that could be taken when studying the effects of face-concealing garments and headgear on speech, and those that will in fact be addressed over the course of this thesis. The next section provides a brief introduction to forensic speech science, and discusses how facewear research can contribute to the field.

2.2 Facewear and forensic speech science

The second part of this chapter introduces the reader to the field of forensic speech science. The various factors which are known to influence speaker recognition performance by expert and lay witnesses are summarised, and most importantly, the contribution of facewear research to forensic phonetics is emphasised.

2.2.1 Forensic speech science in brief

Forensic speech science is a highly interdisciplinary field which applies and extends knowledge, theories and methodologies from (socio)phonetics, (socio)linguistics, speech acoustics, speech technology and signal processing, to practical tasks arising out of the context of police work or the presentation of evidence in court (Jessen, 2008). French & Stevens (2013: 183) estimate that in the United Kingdom alone, forensic speech experts provide witness evidence, or advise in related matters, in approximately 500 cases per year.

The fields of activity in which forensic speech scientists are involved are manifold. They relate, in the broadest sense, to the analysis of audio signals, including those emanating from gun shots, doors banging, machine noise, and the like, and from non-speech human sounds (like coughs and laughter). Most frequently, however, forensic phonetic casework attends explicitly to the analysis of (human) speech. Experts are hence engaged in a wide spectrum of tasks, ranging from audio authentication, audio enhancement, and acoustic reconstruction, to speaker comparison and profiling, speech content analysis, and to some degree forensic linguistic analysis (e.g. in trademark disputes).

This chapter can only sketch an outline of the field and its practical application. To obtain a more comprehensive view, and for further references, the interested reader is referred to articles by French & Stevens (2013), Foulkes & French (2012), Jessen (2008), and Nolan (2001), or to introductory books by Rose (2002) and Hollien

(2002).[7] The goal of this chapter is to frame the contribution of facewear research within forensic speech science. It emphasises that facewear can affect speech on many levels. For this reason it can be argued that facewear effects ought to be accounted for by practitioners when carrying out casework.

On a terminological note, much of the psycholinguistic/cognitive literature refers to the producer of speech stimuli (i.e., the vocalising person) as the 'talker'. In forensic speech science, as well as phonetics and linguistics more generally, the term 'speaker' is more frequently used. Although a semantic differentiation between 'talker' (producer of stimuli) and 'speaker' (of a particular language) seems well-motivated from a linguistic point of view, such a distinction is not intended here. In the following sections, the expression 'speaker' will be retained so as to accord with the wording commonly used in forensic phonetics. In the remainder of the thesis, the terms will be used interchangeably, but preference will be given to 'talker'.

2.2.1.1 Speaker recognition by expert witnesses

The most central purpose of forensic phonetics is the recognition of a person by his/her speech, voice, and language. In this respect, much of the casework centres on 'speaker comparison' and 'speaker profiling'.[8] The key difference between the two is whether or not a speech sample of a suspect in a criminal case is available to the analyst working on the case (Jessen, 2007).

Speaker comparison is the most frequently performed task by forensic speech analysts, accounting for approximately 70% of the casework (French & Stevens,

[7] Additional directed reading, background information, and a collection of case examples, can be found on the websites of the *International Association for Forensic Phonetics and Acoustics* (http://www.iafpa.net/), the *International Journal of Speech, Language and the Law* (https://www.equinoxpub.com/journals/index.php/IJSLL), and *J P French Associates* (http://www.jpfrench.com/) [All accessed: 7th May 2014].

[8] Speaker comparison is also referred to as 'speech comparison' and 'voice comparison' (but see e.g. French *et al.*, 2010, and Rose & Morrison, 2009, for a terminological debate). Speaker profiling is sometimes termed 'voice analysis' (see e.g. Jessen, 2008).

2013: 187). As the name suggests, it involves the comparison of the speech recording of an anonymous speaker who is associated with a crime (hereafter referred to as 'questioned recording') with a speech recording of a known speaker (the suspect). The questioned sample might be the product of a threatening voicemail message, a recorded ransom demand, a fraudulent or hoax call, or a CCTV or covert surveillance recording made by the police or security services (Foulkes & French, 2012).

The goal of forensic speaker comparison is to assist the court in determining the probability of the identity or non-identity of the unknown speaker and the suspect (French & Stevens, 2013: 187).[9] To do so, the analyst inspects the speech samples of both speakers to look for the presence or absence of certain phonetic and linguistic features (see below). Subsequently, the degree of 'similarity' (consistency) of the known and questioned samples, as well as the 'typicality' (distinctiveness) of a particular feature is determined. The former gives the expert an estimate of how compatible the two speech samples are with regard to the evaluated features, and the latter indicates the distribution of each feature in the population of comparable speakers (of the same language, dialect, accent, age group, social status, etc.). The outcome of any speaker recognition activity must always be carefully interpreted on the basis of the analyst's experience and expertise, and in the context of background knowledge from research studies.

A thematically closely-related but less common forensic phonetic task is speaker profiling. Foulkes & French (2012) estimate that *J P French Associates* (the laboratory they are affiliated with) undertakes only about five profiling cases per year. Like speaker comparison, speaker profiling involves the analysis of a recorded sample of speech produced by an anonymous speaker. However, in this case no suspect sample is available for comparison (Jessen, 2007).

The goal of forensic speaker profiling is to gather as much information about the speaker as possible, and to use the resultant profile to help (e.g. the police) to narrow

[9] Expressions like 'identity', 'identification', and 'individualisation' are now generally avoided among forensic practitioners, because experts can never be 100% certain that a person was identified (certainly not from his/her voice alone). Verdicts of this kind are decided on by the trier of fact. That is, the judge or jury in a court case bear sole responsibility for determining the outcome of a trial.

down the list of suspects, or even to find the suspect (Jessen, 2008; Foulkes & French, 2012; French & Stevens, 2013).[10] The analysts are in particular interested in information about the speaker's age, gender, geographical and social background (native/foreign language, dialect, regional accent, sociolect, ethnic origin, etc.), emotional and physical state that would affect speech, or voice disorders and speech/language impediments (stuttering, lisps, etc.).

As mentioned previously, forensic phonetic practitioners performing speaker comparison or profiling analyse speech in respect of a great variety of phonetic and linguistic features (Gold & French, 2011; Foulkes & French, 2012; French & Stevens, 2013). These comprise segmental signal properties (those related to consonants and vowels) and suprasegmental (prosodic) features, like pitch, intonation, tempo, rhythm, and voice quality.[11] Observations of segmental properties are obtained through careful aural-perceptual analysis and fine-grained phonetic transcription and description of the speaker's pronunciations. Experts also account for coarticulation and connected speech processes (e.g. assimilation, elision), and determine spectral (e.g. energy loci of plosive bursts and fricatives, vowel formant frequencies), temporal (e.g. duration, voice onset time), and intensity measures of the segments. In addition, higher-level linguistic properties of the signal and non-linguistic features are commonly taken into consideration. Examples of the former include distinctive morphological, lexical, and syntactical structures, as well as conversational behaviours and discourse markers; examples of the latter are filled pauses, tongue clicking, audible breathing, throat clearing and laughter.

Carrying out this analysis, forensic phoneticians will always reveal differences *and* similarities between two speech samples, even when these originate from the same speaker. The outcome of any speaker recognition task is therefore dependent on the extent to which 'inter-speaker variability' is greater than 'intra-speaker variability'

[10] A specific application of speaker profiling is known as 'Language Analysis for the Determination of Origin'. LADO experts, among other things, assist immigration authorities with establishing the nationality of asylum seekers (Nolan, 2012; Patrick, 2012).

[11] Voice quality can be analysed, for example, by applying Laver's 'Vocal Profile Analysis Scheme', which considers around 38 vocal tract features and settings (Laver, 1980; French & Stevens, 2013). Examples include phonation features (e.g. creaky voice, tremor), vocal tract settings (e.g. nasalisation, pharyngeal constriction), and laryngeal muscular tension.

(Yarmey, 2012). The former term (also known as 'between-speaker variability') refers to the phonetic/linguistic variation in speech produced by different speakers, while the latter expression (also called 'within-speaker variability') specifies the variation in speech samples produced by an individual speaker.

Speaker recognition is further complicated by the fact that within-speaker variation can even be larger than between-speaker variation. Evidently, some voices are more 'distinct' and 'recognisable' than others, which may arise from an unusual combination of relatively rare features. However, some features of the speech of one speaker will always coincide with those of other speakers (Foulkes & Barron, 2000; Nolan, 2001; Dellwo *et al.*, 2007; Watt, 2010; Foulkes & French, 2012). Besides, it is worth keeping in mind that there is no single, invariant, biologically-determined property of the voice that can be used to discriminate between speakers, let alone to establish a person's identity, with absolute certainty (see Nolan's 1983 notion of 'plasticity' of speech).

By and large, the methodologies and procedures applied by practitioners still vary widely across the community (Gold & French, 2011). To date, the most commonly chosen approach is the combination of acoustic and auditory-perceptual analysis; automatic speaker recognition systems are also increasingly adopted in casework. Likewise, until now no overall consensus has been reached on the conventions and conceptual frameworks concerning how to express conclusions (e.g. binary decisions, classical probability scales, 2-step consistency/distinctiveness decision as per the UK Position Statement, likelihood ratios). The reader is referred to Foulkes & French (2012), Gold & French (2011), Jessen (2008), or French & Harrison (2007), for a break-down of methods and conclusion frameworks, and for further references.

2.2.1.2 Speaker recognition by lay witnesses

Forensic experts are also consulted in cases where there is no audio recording of the offender's voice available, but where a witness to a crime has heard a voice and potentially claims to identify the voice (Bull & Clifford, 1984; Hollien & Schwartz,

2000; Wilding *et al.* 2000; Nolan, 2001; Perfect *et al.*, 2002; Yarmey, 2003, 2004; Blatchford & Foulkes, 2006; Eriksson *et al.*, 2010; Yarmey, 2012). Such evidence arises when the (ear)witness could *hear* but not *see* the offender (at least not to the extent that visual identification is possible). Scenarios of this kind emerge, for example, when the offender's voice is heard over the telephone, when the eyes of the witness are covered, when a physical attack takes place from behind or in darkness, and when the criminal's visual (facial) appearance is disguised by facewear during the encounter.

Earwitnesses in above situations typically lack specialised phonetic/linguistic training (hence the use of the terms 'lay' and 'naïve' listeners). When, following the incident, the witnesses are interrogated (e.g. by the police) about the speaker in question, they will therefore give a purely impressionistic description of the perceived voice. On occasion, witnesses claim to have recognised the identity of the speaker. Such statements have led to the development of formal testing of such abilities, which aim to give estimates about the accuracy and reliability of earwitness testimonies.

Most commonly, earwitnesses are asked to participate in a 'voice line-up' (also referred to as 'auditory line-up', 'voice parade', or 'identification parade'). The witness is then exposed to a series of recordings of similar-sounding speakers (which may or may not include the suspect), and is instructed to select the voice believed to be the voice of the offender, or the voice that most resembles it (see e.g. the 'McFarlane guidelines' in Nolan, 2003, or Bull & Clifford, 1984; Foulkes & Barron, 2000; Nolan, 2001, 2003; Yarmey, 2012). Recognition accuracy in these tests can be high when the earwitness was familiar with the speaker prior to the incident. However, research has shown that even recognition of familiar speakers is highly prone to false identification (Foulkes & Barron, 2000). Forensic speech experts generally acknowledge that lay earwitness statements must be treated with great caution by the justice system, and that the probative value of earwitness testimonies is at best questionable (Yarmey, 2012). The factors which are known to influence listeners' judgements of (un)familiar voices are further described in §2.2.2.3.

2.2.1.3 Speech content analysis

In addition to the recognition of speakers by voice/speech, forensic speech scientists are concerned with the (semantic) content of the speech produced by a (known or unknown) speaker. Professional analysts are often assigned the task of providing expert evidence in the form of comprehensive transcriptions or translations. This work relates to cases where the content of a spoken utterance is of particular evidential value, but is difficult to extract without the analyst having professional training and/or access to high-quality audio playback equipment. Examples of factors which are known to complicate speech content analysis include poor quality of the recorded speech material (e.g. noisy, distorted), a foreign language or non-standard accent, and extensive speaker overlap (French & Stevens, 2013).

Sometimes, the speech content in an evidential recording may not only be hard to determine, but may even be ambiguous. This can lead to disagreement between different parties as to what exactly was said in a particular section of the recording. On occasion, two (or more) competing interpretations may be at hand (for example, one provided by the prosecution and one by the defence in a trial). Under such circumstances, forensic speech experts are asked to help to resolve the dispute by carrying out an intensive comparative (aural/acoustic) analysis (French & Stevens, 2013). This is commonly referred to as ‘questioned content’ or ‘disputed utterance’ analysis (French, 1990; French & Harrison, 2006). What should be borne in mind is that even a single highly contentious word can dramatically change the lexical content of an utterance, and therefore possibly the course of an investigation or the direction of argumentation in court. In the most extreme case, differences between words arise only from a single consonant or vowel as a constituent of a minimal pair. Here, the phonological term ‘minimal pair’ refers to words in a language which differ only in one phoneme, such as <**l**ike> and <**b**ike> (see e.g. French, 1990, for a dispute resulting from the near-homophony of <ca**n**> versus <ca**n’t**> in English).

2.2.2 Facewear research in context

The 'ideal' sample of speech for any type of (forensic-)phonetic analysis is one which is sufficiently long, rich in content, not intermixed with background noise, not technically distorted in any way, and which offers a great variety of speaker-specific features. This sample would provide a solid basis for the analyst to express an opinion about speaker 'identity' and speech content. However, forensic reality looks very different. Much of the speech/audio material that practitioners regularly have to deal with is of extremely low quality (and often also quantity). This problem can be so severe that casework inquiries sometimes have to (and should) simply be declined.

The recording or listening conditions in which the voice of the speaker in question was recorded or witnessed are in the majority of cases uncontrolled, and do not match the conditions that the suspect sample is produced or perceived in. For example, suppose that the questioned recording originates from a wiretap hidden in a car, and the suspect sample comes from a police interrogation recorded in a quiet examination room. Or, an offender's voice is witnessed in a noisy and highly stressful situation (e.g. an armed robbery), while the subsequent voice line-up is carried out in a quiet, relatively relaxed environment (e.g. the witness's home).

Forensic practitioners must have a thorough understanding of the numerous factors that are likely to cause a mismatch between samples. They need to be aware that a sample mismatch can negatively impact on the reliability of their own analyses, and that it can cast doubt on lay earwitness statements. The factors known to complicate the estimation of the strength of evidence can be classified as speaker, channel, and listener factors (Alexander *et al.*, 2005). In the following sections, an attempt has been made to place facewear within this framework. It is argued that facewear can be categorised as a speaker, channel, and also a listener factor.[12]

[12] The classifications and terminology differ in the literature. For example, Yarmey (2012) distinguishes between person and system effects. Betancourt & Bahr (2010) differentiate between speaker and mechanical factors. Eriksson *et al.* (2010) report speaker, listener, and situational factors. Alexander *et al.* (2005) talk about speaker, transmission, and system effects. Jessen (2008) discriminates between behavioural and technical factors, and Byrne & Foulkes (2004) between environmental, speaker and technical effects.

2.2.2.1 Facewear as a speaker factor

Speaker factors are those which determine the differences between speech samples obtained from the same speaker or from different speakers. It was explained earlier that these differences can be ascribed to a large number of phonetic and linguistic features, including those which make up a speaker's language, dialect, and accent, and which determine subtle segmental and suprasegmental properties of the speech signal. Moreover, speakers commonly adapt their way of speaking to different occasions and contexts, for example, by changing their speaking style (e.g. read versus spontaneous), speech type (e.g. shouted versus whispered), or register (e.g. formal versus informal). Situation-specific stylistic or paralinguistic variation can also be triggered by distress, health problems, sleep deprivation, alcohol and drug consumption, and voice disguise (for the latter see also §7.2).

In the context of this thesis we add to the list of speaker factors the various types of face- and head-concealing masks and devices. This decision is based on the issues raised in §2.1.2.1 and §2.1.2.2 (first and second research direction). Here, it was exemplified that facewear is likely to actively and passively modify the way that speech is normally produced by the wearer of the mask (i.e., when the face is not disguised). It was further argued that the changes to the speaker's 'natural' speech productions may subsequently alter the acoustic-phonetic properties of the produced speech. The experiments presented in the empirical part of the thesis address these assertions in further depth. For the time being, facewear can be considered as having the potential to increase the mismatch between two samples of speech, that is, to increase the variability in speech samples produced by the same speaker, or in samples produced by different speakers. From a theoretical point of view, it is also conceivable that between-speaker variability might *decrease* in facewear speech, i.e., that two voices become more similar. This notion is taken up in Chapter 6.

2.2.2.2 Facewear as a channel factor

Channel factors refer, firstly, to the factors that can cause qualitative differences between two speech recordings in terms of their technical properties. Most commonly, forensic speech experts have to account for technical interferences and differing transmission characteristics caused by landline (Künzel, 2001), mobile phone (Byrne & Foulkes, 2004; Guillemin & Watson, 2008), and internet telephony (Fecher, 2008), as well as by hardware properties and the quality of audiotapes, digital recorders, wire-tap and other recording devices (Alexander *et al.*, 2005; for details of the 'telephone effect' see §4.3.3). Secondly, channel factors include, in the broadest sense, the environmental conditions in which a recording was made or a voice was witnessed. Examples are ambient noise, the physical distance between speaker and listener or speaker and recording device (which e.g. affects the speech amplitude), or some kind of physical obstacle placed between the speaker and the listener (e.g. other people, a wall in/outside a house).

Against this background, facewear can be classified as a channel factor. It was explained in §2.1.1 that face coverings used to visually conceal the speaker's face can be considered as a 'physical barrier' to the listener. To that extent, facewear acts as a 'passive' element somewhere along the 'speech chain', which may impede the transmission of the acoustic signal. The term 'passive' is used here because the facewear-induced acoustic changes referred to in this context do not involve any action from the speaker. That is, the changes are not connected to the talker's speech productions as such, but refer to the effects of the mask materials themselves (see §2.1.2.2, second research direction). The studies presented later in this chapter, and in the experimental chapters of the thesis, shed some light on the nature of acoustic interference caused by the wearing of facewear.

2.2.2.3 Facewear as a listener factor

Listener factors are those parameters known or expected to impair an expert or a lay witness's performance in aural-perceptual speech and speaker recognition. An important source of misperceptions and transcription errors that arise during expert analysis (especially that of disputed utterances) are so-called priming or expectation effects (French, 1990; Rose, 2002; Fraser *et al.*, 2011; French & Stevens, 2013). Especially in cases where background information on a case is provided, the expert may expect to hear certain words and utterances over others (i.e., they will show a bias towards those that seem more plausible in the broader context of the case). To overcome this bias, analysts need to be aware of the relationship between the processing of 'bottom-up' phonetic/linguistic information (information that is actually present in the signal) and 'top-down' information (information that is supplied by the brain). In other words, they must fully understand the extent to which higher-level linguistic information can interfere with the process of mapping perceptual units onto properties of the acoustic signal (especially where there is phonetic/linguistic ambiguity).

Regarding the limitations of lay speaker recognition, Nolan (2001) sets apart two inherent influencing factors from a large set of contingent factors. Inherent factors are the performance of the human perceptual, storage, and retrieval mechanisms, and the overlap of different voices in terms of their phonetic/linguistic properties (see §2.2.1.1). Contingent factors include the listener's age, gender, hearing ability, familiarity with the speaker or the speaker's language/dialect/accent, attentional and cognitive capacities, stress, health, emotional status, general expectations and individual skill sets. Furthermore, a lay witness's performance in recalling a particular voice can be influenced by the retention interval (time elapsed between initial exposure to the voice and recall from memory), the distance between speaker and listener, the number of times the voice was heard, the type of voice exposure (active = speaker and listener interacted, passive = listener only overheard the speaker), or the length of the perceived utterance (Ladefoged, 1978; Schiller *et al.*, 1997; Schiller & Köster, 1998; Nolan, 2001; Yarmey, 2003, 2004; Blatchford & Foulkes, 2006; Eriksson *et al.*, 2010; Watt, 2010; Yarmey, 2012).

Once again, facewear can be added to the list of known listener factors. Earlier in this chapter it was noted that there is a possibility that facewear, on one level or another, has an impact on the perceiver of facewear speech. In §2.1.2.3 (third research direction) it was also noted that auditory(-visual) speech processing might be impaired when the speaker (or listener) is wearing facewear. The following sections present previous research which addressed related issues. The effect of facewear speech on the listener is also investigated in the experiments discussed in later chapters. For the time being, we can acknowledge that the quality of aural-perceptual analyses by experts, and the reliability of earwitness testimony by lay people, are likely to be further compromised by facewear.

2.3 Previous research on facewear

The third part of this chapter gives an account of preceding research on the subject matter. To start off, studies which have looked at the influence of facewear on speech explicitly from a forensic point of view are surveyed (especially Llamas *et al.*, 2008). The chapter concludes with a review of work that has examined the effects of various types of face and head coverings on speaking and listening more generally.

2.3.1 Llamas, Harrison, Donnelly & Watt (2008)

The article by Llamas *et al.* (2008) reports on two experiments. The first experiment tested the effects of different types of forensically-relevant mouth and face coverings on speech perception. The second experiment addressed the effects of facewear on the acoustic speech signal. The objective of the former experiment was to ascertain whether speech intelligibility is adversely affected when the talker's face is disguised by a face covering, and if so, to what extent this effect is the result of disruption to/absence of *visual* speech cues from the talker's face, or the consequence of disruption to/absence of *auditory* speech cues.

Thirteen mostly native English speakers participated in the experiment (10 females, 3 males; age range = 18–37 years). The speech material consisted of 40 mostly monosyllabic English words with predominantly CVC syllable structure, which were chosen to exemplify a range of consonants in onset (/p t k s ʃ z f v h θ ð m n/) and coda (/p t k s ʃ θ t ʃ dʒ m n ŋ ts/) positions. The syllable nucleus was one of four vowels (/i ɪ a ɔ/). All target words were embedded in the standardised carrier sentence *Say <target word> again*.

The resultant sentence list was read in control (undisguised) condition by a female (aged 23) and a male (aged 25) native Scottish English speaker. Both talkers then repeated the list three times, each time wearing a different type of facewear. The face

coverings tested in this study were a balaclava (without mouth hole), a *niqāb*, and a surgical mask.

Video recordings of the talker's head and shoulders (frontal view) were made of all reading sessions. Two types of stimuli were produced from the recordings. They are henceforth referred to as 'auditory-visual' (AV) and 'auditory-only' (AO). In the AV condition, participants saw the moving image and heard the soundtrack of the videos. In the AO condition, no image was presented, i.e., participants were exposed to the talker's voice only. The participants' task was to write down the target words they perceived on answer sheets provided to them (word familiarity was controlled for). In total, each participant was exposed to 640 test utterances: 2 talkers x 40 target words x 2 modalities x 1+3 facewear conditions (control + 3 types of facewear).

Once data collection was completed, and spelling and vowel mistakes had been eliminated, the participants' responses were rated for consonant recognition errors, which emerged when a target word was misheard (e.g. <**th**in> as <**f**in>, <si**p**> as <si**t**>, <pi**p**> as <pi**ck**>, <ki**n**> as <ki**ng**>, etc.). Llamas and colleagues found that only a small number of misperceptions occurred across facewear conditions. Specifically, only about 2% of the 8,320 responses collected in total (640 stimuli x 13 participants) were incorrect. According to the authors, this suggests that the participants had correctly identified the target words with a high degree of reliability.

As expected, a higher rate of 'consistent' misperceptions (i.e., those reported by three or more participants) was registered in the AO than in the AV condition. Given that the audio signal presented to participants was identical in both conditions, the authors inferred from this result that the visual speech information encoded in the talker's face plays a vital role during consonant recognition. More importantly, this appeared to be the case even when the face was concealed by a face mask. Among the three tested guises, the overall highest proportion of consistent misperceptions was produced in the balaclava condition.

Furthermore, the authors observed that only a small number of confusion types accounted for the majority of listening errors. Most common were the confusion of stops with (mainly homorganic) fricatives (especially /t/ with /θ/), difficulties in correctly identifying the place of articulation and voicing feature of stops, as well as

misjudgements of the place of articulation of fricatives (especially /f/ with /θ/) and nasals (especially /n/ with /ŋ/).

In the second experiment carried out by Llamas *et al.* (2008) an attempt was made to isolate the acoustic effects of facewear (i.e., those that were not related to changes to the talker's speech productions and acoustics). To do so, the *transmission loss* characteristics of a range of fabrics were measured. Transmission loss (TL) is described by the authors as the property of the material that relates to its frequency-dependent sound attenuation characteristics. Llamas *et al.* further explain that different energy loss mechanisms of different materials will result in a greater or lesser degree of acoustic attenuation in different parts of the spectrum. The fabrics examined in this study were woven polyester (the material of the *niqāb* tested in the listening test), knitted acrylic (the tested balaclava), pleated paper (the tested surgical mask), woven cotton (a handkerchief), knitted wool/acrylic mix (a woollen scarf), knitted polyester (a fleece scarf), 1-denier sheer nylon (stockings), and a woven 'acoustically-transparent' cover fabric (used to conceal loudspeakers, absorbers and diffusers for aesthetic reasons).

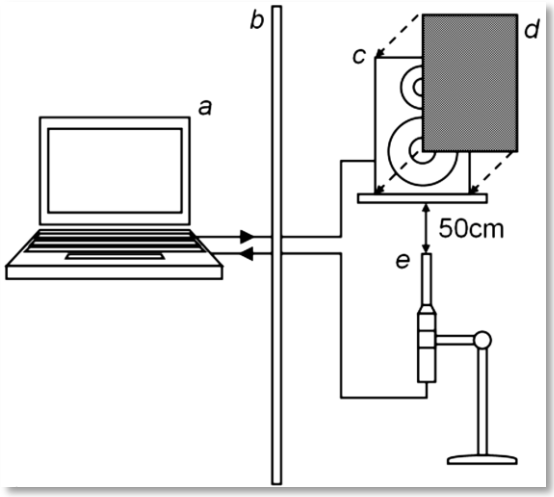

Figure 2.1. Experimental set-up employed during the transmission loss experiment reported in Llamas *et al.* (2008). 'a' = control PC; 'b' = soundproofed partition wall; 'c' = loudspeaker; 'd' = fabric sample; 'e' = microphone. Reproduced from Llamas *et al.* (2008: 93).

Figure 2.1 illustrates the experimental set-up used to assess the TL characteristics of the aforementioned fabrics. Samples of each material were interposed between a loudspeaker (acting as the sound source) and a microphone in an acoustically-treated recording booth. The authors employed the Maximum Length Sequence method (Rife & Vanderkooy, 1989) to measure the impulse response of the microphone-loudspeaker system both with and without fabrics intervening. The TL of the fabrics corresponded to the difference in frequency response between the control (no fabric) and the fabric conditions.

Contrary to expectations, the differences between the TL obtained in the control condition and the TL in each of the fabric conditions were only minor. The only exception was the surgical mask (and the experimenter's body placed between loudspeaker and microphone, which was introduced as an additional extreme condition). As Figure 2.2 shows, the TL caused by the surgical mask deviated from the TL in the control condition most notably between 2.5kHz and 12.5kHz, and between 14kHz and 24kHz (upper cut-off frequency).

Figure 2.2 further reveals that transmission *gain* (negative TL) occurred on occasion, which means that particular frequencies had greater amplitude *after* the signal was filtered through the fabric (see e.g. the surgical mask at 1–2kHz). The authors suggest that transmission gain ought to be interpreted in the light of the fabrics acting as an interactive component of the loudspeaker-microphone system, rather than a mere blockage or attenuation element.

Llamas *et al.* (2008) acknowledge that the set-up used in the TL experiment does not adequately reflect natural speech production through facewear fabrics (e.g. lack of airstream and radiation factor for a close-fitting mask). But despite its limitations, the study offers valuable first insights into the speech transmission characteristics of a range of fabrics. With a view to the acoustic experiments presented in later chapters of this thesis, it is worth keeping in mind in particular the finding that thicker, heavier materials do not automatically cause greater TL than thinner, lighter ones, but that the relationship between different materials and their sound absorption characteristics is a rather complex one.

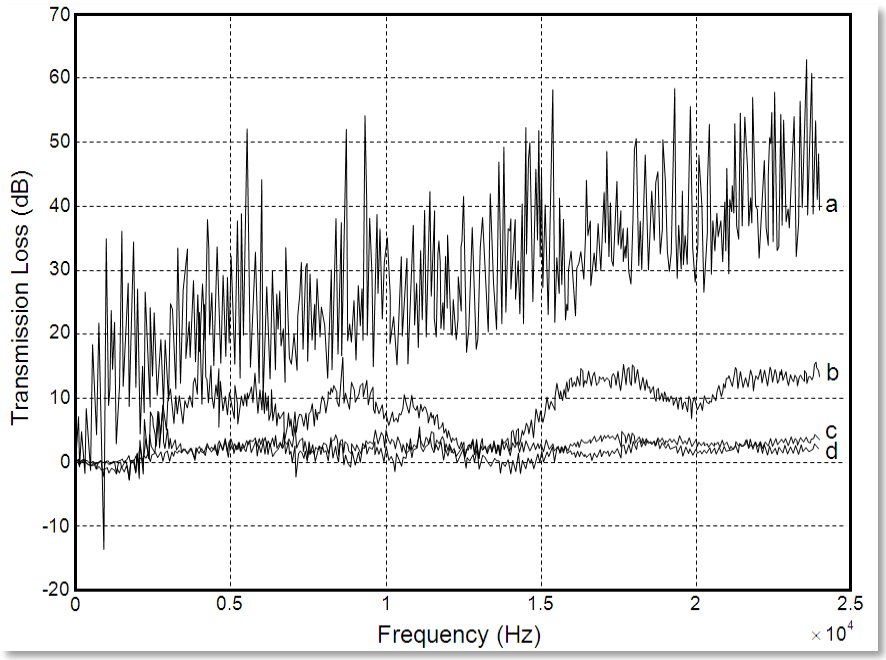

Figure 2.2. Transmission loss differences between the frequency response curves for the control condition and the four test conditions. 'a' = human body; 'b' = surgical mask; 'c' = woollen scarf; 'd' = *niqāb*. The zero line denotes parity of the frequency response in a test condition with the response in the control condition at any point between 0–24kHz. Reproduced from Llamas *et al.* (2008: 95).

In conclusion, Llamas *et al.* (2008) suggest that the detrimental effect of facewear on speech intelligibility must derive principally from the disruption to/absence of visual speech cues available to participants, and from the auditory consequences of interference of the facewear with speech articulation, but probably to a lesser degree from the auditory consequences of acoustic transmission loss induced by the facewear material itself.

On a final note, the results reported by Llamas *et al.* (2008) conform with preceding TL studies which adopted similar experimental designs. For example, Nute & Slater (1973) examined the sound transmission characteristics of 44 woven fabrics over a broad range of frequencies. They found that sound absorption was mostly dependent on the weight, thickness, and cover (density/porosity) of the fabric, and that TL was more prominent at higher frequencies. The latter may be to do with the short wavelengths of high frequencies, which are more readily affected by flow-resistance in the fabrics. Aso & Kinoshita (1963a, 1963b, 1964) also report that the degree of porosity in woven fabrics needs to be considered relative to the sound levels when

assessing TL. More recently, Noy (2003) measured the frequency response of a loudspeaker-microphone system with and without fabrics interposed, and observed that all tested materials, including an 'acoustically transparent' loudspeaker cover, reduced signal amplitude, most particularly above 10kHz.

2.3.2 Other forensically-motivated work

To date there exist, to the best of the author's knowledge, only two published forensically-motivated studies other than Llamas *et al.* (2008) which investigated facewear effects on speech. In 2008, Zhang and Tan experimentally determined the effect of voice disguise on the performance of a forensic automatic speaker recognition system developed by the same authors. The ten types of voice disguise tested in the study were foreign accent, whisper, raised/lowered pitch, fast/slow speech, pinched nostrils, use of (bite block) objects (pencil, chewing gum), and most importantly, facial masking. The facewear used here was a surgical mask composed of relatively thick cotton. It was rather loose-fitting, and covered the talker's mouth and parts of the nose (Cuiling Zhang, personal communication, 12th January 2011).

The test material consisted of speech recordings of 20 male native Standard Chinese speakers in their early 20s. The first part of the study established the system's performance with *undisguised* voices. Speech samples from all talkers were added to a pre-existing database of 2,000 talkers, and then used as test samples for automatic speaker identification and verification. The results showed that nearly all talkers were correctly recognised, which confirmed the good system performance reported in earlier studies by the authors.

Next, the same test procedure was applied for the disguised voices, and each disguised voice was compared with each undisguised voice in the database. It was found that system performance was greatly reduced when voice disguise was introduced. The magnitude of this decrease varied with the type of disguise. Most interestingly, the correct recognition rate was lowest in the surgical mask condition

(0%), along with whisper (0%) and raised pitch (10%). By comparison, the lowered pitch, pinched nostril, and pencil conditions led to 45%, 55%, and 65% correct recognition, respectively. The changes to speech tempo and accent yielded >85% accuracy. Altogether, the study demonstrated a highly disadvantageous effect of facial masking on automatic speaker recognition (for this particular type of facewear and system).

The second forensically-motivated study to consider the masking of a person's face as a potential influencing factor on speaker recognition was carried out in 2011 by Heath and Moore. The scope of their work was, however, a very different one. Their study focused on the interaction of auditory and visual information relating to a talker during recall of the talker's voice by a (human) listener (voice memory).

This line of research has its origin in a phenomenon called the 'verbal overshadowing effect' (Schooler & Engstler-Schooler, 1990). The effect implies that describing a particular stimulus verbally can interfere with the memory (and subsequent recognition) of that stimulus. For example, the verbal description of a voice can reduce recognition accuracy in a following voice line-up (Perfect *et al.*, 2002). This reflects findings from the visual domain, where the description of a face has been found to impair face recognition (Dodson *et al.*, 1997). Interestingly, studies have shown that even the mere presentation of a *face* together with the voice during initial exposure can compromise voice memory. This phenomenon has been termed the 'face overshadowing effect' (Cook & Wilding, 1997, 2001).

Heath & Moore (2011) examined the face overshadowing effect and included two novel variables in their experimental design, namely 'facial concealment' (with a balaclava) and 'emotionality of vocal tone' (angry versus neutral). The researchers hypothesised that the magnitude of the face overshadowing effect would be reduced in the balaclava condition (compared to the control condition, where the talker's face was unconcealed) on the grounds that facial disguise of the talker would increase the listener's focus on the voice, and limit interference with the visual (facial) stimulus.

During the presentation phase of the study, the first participant group listened to the voices of six unfamiliar talkers consecutively; the second group was exposed to the same voices and additionally to the talkers' faces; the third group listened to the

voices and watched the faces, but this time the faces were each disguised with a balaclava (one with a mouth hole). In the second stage of the experiment, the participants were auditorily presented with four target voices and four distractor voices (foils). Their task was to judge whether or not they had heard a particular voice before.

The results showed that the participants on average recalled only 2.14 of the four target voices (angry voices were by trend better recognised). Contrary to predictions, the presentation of a balaclava-concealed face *did* give rise to the face overshadowing effect. The mean recognition rates were lower when the observers were additionally exposed to the talkers' faces (compared to the voices only), irrespective of whether or not the faces were obscured by a balaclava. The study hence reinforced that voice memory is impaired when a voice is presented simultaneously with the talker's face at encoding. The new finding is that this appears to be the case even when the face is visually disguised by means of facewear (for further discussion of Heath & Moore's results see §7.3).

2.3.3 Thinking outside the (forensic) box

Research on the effects of facewear on speech from a forensic perspective is scarce. However, when thinking outside the box and consulting the literature, one finds that several studies have in fact previously addressed questions concerning interpersonal communication when facewear of some sort is involved in the communication process between two (or more) interlocutors. In this regard, questions relating to the impact of facewear on the acoustic speech signal are as much of concern as those dealing with speech intelligibility on the part of both the talker and listener.

In an educational context, Coniam (2005) examined the extent to which the scores awarded to students in an oral examination are influenced by audibility and comprehension problems encountered when the students and teachers are wearing surgical masks. The study was performed in 2003 in Hong Kong during the outbreak

of the Severe Acute Respiratory Syndrome (SARS). At that time, many citizens wore surgical masks in an attempt to protect themselves from being infected by the virus. Participants in the study took part in a mock oral English language test. This was completed under the condition that all test takers and examiners were either wearing a face mask, or not. In the mask condition, the teachers and students hence had to interact with each other with some of their facial expressions and visual speech cues removed.

Coniam found that the students' performance in the test was not markedly reduced in the mask condition (except for lower pronunciation marks). Interestingly enough, the students indicated in post-test interviews and questionnaires that they had adopted certain compensatory strategies to counteract the constraints imposed by wearing the surgical mask. Examples given include slower speaking rate, increased speaking volume, clearer articulation, enhanced use of 'body language', and more eye contact with the examiners. Several students also expressed their concern about not being able to see the examiners' facial expressions.

Furthermore, studies of facewear effects on speech often target the types of face masks worn for occupational purposes. Of particular interest in this respect is the headborne 'personal protective equipment' (commonly abbreviated as 'PPE') worn in industrial, military, and medical environments to shield the wearer from breathing in high levels of smoke, fumes, gas, or vapour, or from inhaling and spreading potentially dangerous airborne microorganisms. The most extensively studied masks are various models of air-purifying respirators, surgical masks, and hearing protection devices.

The primary goal of many studies is design- and functionality-oriented. Usability tests are regularly run, with the aim of checking the gear for suitability, effectiveness, ergonomic factors like comfort and fit, and – most interestingly to the present context – communicative efficiency (see e.g. Howell & Martin, 1975; Abeysekera & Shahnavaz, 1987; Wilde & Humes, 1990; Tubbs, 1995; Eck & Vannier, 1997; Pääkkönen *et al.*, 2000; Wijngaarden & Rots, 2001; Caretti & Strickler, 2003; Tufts & Frank, 2003; Wagoner *et al.*, 2007; Mendel *et al.*, 2008; Roberge, 2008; Wittum *et al.*, 2013). The urgency of carrying out these tests is accentuated by the fact that the

masks are often worn for a considerable length of time, and that equipment failure can compromise work quality and safety (especially in professional environments which exhibit high levels of background noise, such as factories, building sites, medical and healthcare institutions, or in military aviation and firefighting).

It may not come as a surprise to the reader that habitual wearers associate the wearing of PPE with problematic hearing and compromised verbal communication. Reports of this kind are often anecdotal, and stem from surveys or questionnaires among the workforce. For example, Bensel *et al.* (1987) report that the mask and hood of a standard chemical protective clothing system worn by army personnel interferes with the wearer's ability to understand spoken words and to be understood when speaking. Coyne *et al.* (1998), Bishop *et al.* (1989), and Howell & Martin (1975) report the use of alternative behaviours to compensate for difficult speech communication, such as hand signals, or firefighters pulling away the respirator facepiece of their breathing apparatus before calling out to one another. Furthermore, Salazar *et al.* (2001: 238) mention a survey which revealed that a group of workers on a hazardous waste site perceive the inability to 'hear and be heard' as one of the most negative aspects of respirator use. A survey among hospital staff, cited in Roberge (2008), showed that nearly half the respondents associated the wearing of respiratory equipment with communication difficulties. Similarly, Wittum *et al.* (2013) point out that surgeons regularly complain about reduced intelligibility in operating rooms.[13]

Scientific research into impediments to speech production brought about by PPE and other face masks typically investigates speech intelligibility by means of perception tests, where listeners are presented with standardised speech material (often the Modified Rhyme Test; see e.g. House *et al.*, 1965; Sommer, 1976; Caretti & Strickler, 2003; Kapoor, 2012; Radonovich *et al.*, 2010) which has been produced in 'mask' and 'no mask' conditions. For example, Bishop *et al.* (1989) assessed the

[13] Some of these communication problems can now be mitigated by means of speaking membranes, voice amplifiers, earplugs with integrated miniature earphones, and other advanced technologies, and/or by standardising usability requirements (see e.g. Goldfrank & Liverman, 2008, and Coyne & Barker, 2010, for recommendations given by the U.S. Institute of Medicine and Department of Homeland Security).

communicative efficiency of different types of chemical-biological warfare masks worn by field personnel, and found that the distance between the interlocutors was crucial to word intelligibility. Eck & Vannier (1997) observed that various respirator interfaces worn by healthcare workers impair verbal communication. They ascribed the problem to the reduced intelligibility and volume of the produced speech. Coyne *et al.* (1998) report that the larger the distance between talker and listener, and the less semantic context is offered (single words as opposed to predictable sentences), the lower the comprehension of spoken words and phrases produced through a respirator in noise. Abeysekera & Shahnavaz (1987) found that half- and full-face dust respirators only marginally interfered with speech transmission. However, the authors point out that a moderate degree of interference with intelligibility caused by a mask may, under some circumstances, be more dangerous in a workplace (as it can lead to a wrong action) than the complete loss of a message (which may lead to no action).

More recently, Radonovich *et al.* (2010) quantified the effects of various disposable and reusable respirators and surgical masks worn by healthcare workers in an actual hospital-based environment and in a simulated workplace. They found that intelligibility was dependent both upon the type of mask and the environment the mask was used in. For example, speech intelligibility decreased quite substantially for some models (by 10–17%), but less so for others. In the 'no mask' condition, intelligibility approached 100% in the simulated environment, but only 90% in the authentic hospital setting (possibly owing to room reverberations and distracting noise from machines). These findings accord with those of Wittum *et al.* (2013), who determined the degree of degradation of speech communication caused by two types of surgical masks (with and without blood shield attached) worn by anaesthesiologists and surgeons in operating rooms. Here, 21 listeners participated in a speech-in-noise test in which they were to repeat particular words embedded in low and high predictability sentences. The results showed that performance was generally poorer on low predictability sentences (see Coyne *et al.*, 1998), and that recognition accuracy was highest in the 'no mask' condition (48.5%) and lower in the two mask conditions (with blood shield = 33.1%, without blood shield = 20.9%).

Apart from studying the perceptual effects of PPE, researchers have looked at the impact of the devices on the acoustic speech signal. In one of the earliest accounts, Morrow (1947) reports formant shifts for speech produced behind a gas mask, noise shield, and oxygen mask. Later, Bond *et al.* (1989) analysed acoustic-phonetic characteristics of speech produced in noise through an oxygen mask integrated in an Air Force standard flight helmet, and found an increase in vowel and word duration, fundamental frequency (F0), and total energy. Vojnović & Mijić (1997) analysed long-term spectra of speech produced behind an oxygen mask worn with a flight helmet. They observed that the speech was only marginally affected by the presence of the mask in the 100–800Hz frequency band, but found evidence for spectral changes above 800Hz (maximum relative attenuation measured at 2.5kHz).

Stanton *et al.* (1988) measured a wide range of acoustic-phonetic properties of English phonemes (e.g. F0, formants, duration, spectral tilt and centre of gravity), which were produced in three conditions: 'normal', 'loud' (10dB above normal), and 'Lombard' (where noise was played back through headphones during speech production). The five talkers were each wearing complete flight headgear (helmet/oxygen mask) while seated in a fighter cockpit simulator. The results revealed that lower (0–0.5kHz) and higher (4–8kHz) frequency bands of vowels and liquids produced in the loud and Lombard conditions lost energy relative to the mid-frequency region (0.5–4kHz). The authors further registered an overall shift of energy towards the higher frequency bands in voiceless fricatives and stops. In other words, more energy was now concentrated in higher frequencies (at 4–8kHz). The phenomenon that energy in certain frequency regions increases at the expense of energy in other regions was termed 'energy migration' by the authors (Stanton *et al.*, 1988: 322).

Finally, some studies have specifically examined the extent to which hearing, speech intelligibility, and speech acoustics are affected by equipment which covers the talker's ears, such as communication headsets, helmets fitted with protective ear cups, earplugs, earmuffs, and other hearing protection devices. The wearing of such gear is vital in very noisy workplaces to mitigate noise-induced hearing loss (Wagoner *et al.*, 2007). For this reason, a lot of effort is put into evaluating the efficiency of the devices, often by means of audiometric tests.

One of the main problems in connection with hearing protectors is that they not only provide attenuation of the (unwanted) noise, but also filter out portions of the (wanted) speech signal. Hence, most relevant to the present context are studies which consider high noise exposure and low speech intelligibility caused by high levels of background noise as connected problems, and examine to what extent the two competing goals – reducing noise exposure while maintaining speech communication – are met. This was addressed e.g. by Wagoner *et al.* (2007), who were testing different models of earplugs, or by Wijngaarden & Rots (2001), who were experimenting with earplugs and helmets worn by Chinook helicopter aircrews (see also Howell & Martin, 1975; Abel *et al.*, 1980; Wilde & Humes, 1990; Pääkkönen *et al.*, 2000).

Other researchers have looked at speech perception by listeners whose ears are *not* covered, but who are exposed to speech produced by talkers whose ears *are* covered while speaking. Speech produced while the talker's ears are occluded can lead to impaired auditory feedback of the talker's own voice, and consequently to articulatory/acoustic changes to his/her speech. Studies have shown that 'earwear speech' will be less intelligible to the listener than the same talker's control (no earwear) speech (e.g. Kryter, 1946; Howell & Martin, 1975; Martin *et al.*, 1976; Tufts & Frank; 2003). In a thorough examination of this effect, Tufts & Frank (2003) obtained intelligibility ratings as well as intensity and spectral measures of speech produced in quiet and noisy environments while the talkers had one of two types of earplugs inserted into their ear canals. In the quiet environment, the acoustic and perceptual properties of the signal were similar between the 'earplug' and 'no earplug' condition. For example, the talkers only marginally lowered the level of their voice (by ~0.6dB) when wearing earplugs. In noise, on the other hand, speech intelligibility was compromised, and the spectral properties of the signal were modified when the talkers wore earplugs (for example, there was more high-frequency energy in the spectrum). The talkers raised the level of their voice in both conditions, as expected, but intensity was relatively lower (by ~4–11dB) in the 'earplug' than 'no earplug' condition. The authors suggest that the reduced intensity of the voice may be the consequence of the earwear attenuating the perceived

ambient noise level, or of an enhancement of bone-conduction hearing at frequencies below 2kHz (which would result in the talker's voice appearing louder to the talker).

In sum, Chapter 2 presented relevant literature, placed the current research in the field of forensic speech science, and outlined the research directions taken in the thesis. Before moving on to presenting the empirical research, Chapter 3 describes the design of the database which provided the speech material for all experiments.

3 The ‘Audio-Visual Face Cover’ corpus

3.1 Corpus design

The present chapter describes the design of a speech database consisting of carefully controlled, high-quality audio and video recordings of talkers whose faces were concealed by a comparatively large variety of forensically-relevant face and head coverings at the time the speech was produced. The corpus provides the basic dataset utilised in all experiments presented in the empirical chapters of the thesis. It is hereafter referred to as the 'Audio-Visual Face Cover' (AVFC) corpus. This chapter introduces the talkers and recorded speech material, the types of facewear that were incorporated in the study, as well as the recording set-up and post-processing techniques (as previously outlined in Fecher, 2012).

3.1.1 Talkers

The 'Audio-Visual Face Cover' corpus (henceforth AVFC corpus) consists of recordings of ten talkers, five females and five males. Their ages ranged from 21 to 36 years ($\bar{x}$ = 26.5, *SD* = 5.7). No participant reported a history of impaired speech, hearing or vision. All were native English speakers who speak with a Southern Standard British English accent. Furthermore, all had a linguistics and/or phonetics background, and held a degree in linguistics (from B.A. to Ph.D. level) at the date of the recordings. Lastly, all talkers had had previous training in the use of the International Phonetic Alphabet (IPA). This enabled them to produce the stimuli presented to them in IPA characters reliably and consistently.

No participant reported prior experience of wearing any type of facewear for recreational, occupational or religious/cultural purposes on a regular basis (as indicated by the questionnaire shown in Appendix C.1, which was completed by each participant). Given the variety of facewear tested, it seemed more feasible to recruit talkers with limited experience of wearing facewear. This factor was controlled for because people who routinely wear a face covering (e.g. surgical nurses and doctors, or wearers of the *niqāb*) might compensate for known

disadvantageous (auditory) effects more extensively (e.g. by speaking more loudly or using more exaggerated articulation) than those who do not. It was decided to record more than just one or two talkers in order to be able to compensate at least to some degree for intra-talker variation in the speech material used for the studies presented in Chapters 4 to 6, and also to be able to generalise the experimental findings.

All talkers were staff and students recruited from the Department of Language and Linguistic Science, University of York, United Kingdom. They were paid for their participation. Prior to taking part they were informed about the procedure both in written and verbal form so that they could grant their informed consent to participate. The data collection was approved by the University of York Humanities and Social Sciences Ethics Committee (for accompanying documentation see Appendices B.1 and B.2).

3.1.2 Facewear

As defined in Chapter 1, the term ‘facewear’ is introduced in this thesis to refer to the various types of face-concealing garments and headgear that are worn in everyday communication situations, as well as in the context of crimes and situations of public disorder. Figure 3.1 shows profile and half-profile photographs of one of the male talkers recorded for the AVFC corpus while wearing the following face coverings:

1. motorcycle crash helmet (visor kept raised)
2. balaclava (without a mouth hole)
3. strip of adhesive tape across the mouth/cheeks
4. balaclava (with a mouth hole)
5. *niqāb* (full-face veil)
6. surgical mask
7. hoodie (hooded sweatshirt) and scarf combination
8. full-head rubber mask

This selection of facewear once again illustrates that the choice of masks for the present study was not only motivated by their direct forensic relevance, but was also targeted at ordinary spoken communication situations out of which a forensic case could potentially arise (see §1.1.1).

Figure 3.1. Profile and half-profile images showing one of the male talkers recorded for the 'Audio-Visual Face Cover' (AVFC) corpus in the control (no facewear) condition (upper left) and while wearing each of eight types of facewear. The selection criteria for the facewear were (potential) forensic relevance, the region of the talker's face that was occluded by the mask, and the facewear material.

The second selection criterion besides (potential) forensic relevance was to do with the regions of the talker's face that were obscured by a particular face covering. The intention was to include a variety of facewear which would cover different parts of the talker's face. The images in Figure 3.1 show that one facial disguise only covered the mouth (tape), while other masks additionally occluded the nose (surgical mask) and ears (motorcycle helmet, hoodie/scarf, balaclavas). In some cases nearly the

entire head and face (except for the eyes) was concealed (rubber mask, *niqāb*). This characteristic of the facewear will be of particular relevance in the auditory-visual speech perception experiments presented in Chapter 5. It will be investigated whether listeners can still extract visual speech information from a talker's articulating face when some portions of it are no longer visible to them.

Lastly, the third selection criterion for the facewear in this study concerned the material which covered the talker's head/face, and in particular, the mouth and nose region. Table 3.1 lists the materials that the facewear was manufactured from.

facewear	facewear material
motorcycle crash helmet	lightweight thermo composite shell, form-fitted contoured cheek pads, absorbent inner cloth, ventilator system near mouth, visor kept raised
balaclava (no mouth hole)	100% cotton double-knitted fabric
strip of adhesive tape	5cm wide, flexible, microporous surgical tape; gentle, hypoallergenic adhesive; inner surface slightly loosened from the talker's lips
balaclava (mouth hole)	acrylic double-knitted fabric, with an extra fleece lining
***niqāb* (full-face veil)**	1-layer *niqāb* (satin headband), worn on top of a cotton/polyester *hijab* ; lightweight, grid-like polyester chiffon from eyes downwards
surgical mask	pleated 3-layer non-woven paper-like fabric, elastic ear loops and nose piece, talker's mouth and nose loosely covered
hoodie/scarf combination	100% cotton hooded sweatshirt, 100% cotton bandana (kerchief) tightly but comfortably wrapped around the talker's mouth and nose
rubber mask	full-head soft latex mask, small holes for each eye, hole in the mouth region (5cm wide, 1cm high)

Table 3.1. Facewear material of each of the eight types of face-concealing garments and headgear worn by all talkers who were recorded for the AVFC corpus. The face coverings were selected so as to represent a fairly large variety of materials.

The mask material was considered a crucial factor in the present context. On the basis of the literature discussed in §2.3 it was hypothesised that some materials would attenuate sound energy in different frequency bands and to different degrees. For example, the thick outer shell of the motorcycle helmet, or the double-knitted fabric of the balaclava, were expected to absorb sound energy to a much greater extent than the thin textiles of the surgical mask or the *niqāb* (see also Chapter 4).

All talkers wore the same facewear during the individual recording sessions. Naturally, these items fit the talkers to varying degrees, depending on the size and shape of their heads. The facewear may for this reason have perturbed speech articulation to a larger extent for some talkers than for others. This factor ought to be taken into account when interpreting the results of the speech acoustic and perception studies presented in later chapters.

3.1.3 Speech material

Prior to reciting the main target stimuli, each talker read aloud the reading passage 'The boy who cried wolf' (see Deterding, 2006; full text shown in Appendix C.2). The aim was to obtain phonetically-controlled reference material for each talker. Having the participants read the text was furthermore intended to reduce their stress levels at the outset of the recording session, i.e., for them to accustom themselves to the experimental set-up during the recordings. The recording sessions took place in a large, professional recording studio (see §3.1.5).

The list of target stimuli was specifically designed for the purposes of this corpus and the intended acoustic and perception experiments. It consisted of phonetically-controlled /C_1ɑ:C_2/ syllables, which were embedded phrase-finally in the carrier sentence *He said [stimulus]*. The carrier phrase was presented to participants in ordinary orthography, while the target syllables were displayed as IPA characters so as to avoid ambiguity of pronunciation.

The target syllables consisted of two tokens each of 18 consonants in two syllable positions. The nucleus was always the open back vowel /ɑ:/. 18 English consonants, namely /p t k b d g f s ʃ θ v z ʒ ð m n ŋ h/, occurred twice in onset and coda position, respectively (for exceptions see below). Consonants were each time spoken in a different phonetic environment, i.e., with a different 'filler consonant' (which was not the target). This was to compensate for connected speech processes, such as anticipatory or carryover nasal coarticulation, that might occur. Finally, English

phonotactic constraints were observed: /h/ in coda and /ŋ/ in onset position were excluded. Hence, these two phonemes occurred only once apiece, making a total of 64 syllables per list.

All stimuli were logatoms (nonsense words), so as to prevent top-down processing caused by higher-level factors such as lexical predictability or contextual plausibility from biasing recognition performance in subsequent perception experiments (see also §5.2.1.2; following e.g. Ganong, 1980; Bernstein & Auer, 1996; Rönnberg *et al.*, 1998; Bernstein *et al.*, 2000; Massaro, 2001; Cutler *et al.*, 2004; Sheffert & Olson, 2004; Lidestam & Beskow, 2006; Stephens & Holt, 2010). To eliminate the occurrence of common real words in the stimuli set, all tokens were checked by three native English speakers. Existing one-syllable words were replaced by changing the filler consonant. Altogether, this procedure resulted in the following list of target syllables: [pɑːʒ, gɑːp, dɑːm, pɑːz, zɑːt, tɑːv, fɑːʃ, pɑːb, tɑːs, fɑːb, tɑːg, fɑːz, hɑːs, fɑːn, dɑːp, ʒɑːf, pɑːʃ, tɑːb, fɑːŋ, kɑːʃ, ðɑːt, pɑːn, bɑːp, tɑːʃ, sɑːf, fɑːð, ʃɑːf, pɑːg, ʃɑːs, fɑːθ, mɑːp, fɑːf, nɑːp, ʃɑːb, θɑːp, tɑːf, sɑːk, fɑːd, mɑːf, fɑːʒ, tɑːθ, dɑːf, tɑːk, vɑːf, bɑːʃ, fɑːg, gɑːf, tɑːd, zɑːf, ʃɑːm, ðɑːf, sɑːt, pɑːŋ, hɑːb, θɑːf, gɑːk, vɑːt, kɑːg, fɑːp, tɑːð, fɑːv, ðɑːp, ʒɑːt, nɑːf].[14]

In sum, each of the ten talkers read a list of 612 stimuli sentences, resulting in a total of 6,120 recorded utterances: 10 talkers x 18 consonants x 2 syllable positions x 2 repetitions (excluding phonotactically-invalid syllables) x 1+8 facewear conditions (control + 8 types of facewear).

[14] Some of the test syllables are real words because lexical gaps with English monosyllables are hard to find. Given that they are low-frequency words they were kept in the stimulus list.

3.1.4 Prompting method

The order of the 64 syllables in the stimulus list was randomised nine times, thus obtaining lists 1–9. After each talker had completed a brief practice run, they read aloud all nine lists in random order, following a predefined recording protocol (see Appendix C.3). One list was thus read in the control (no facewear) condition and the remaining eight lists with the talker each time wearing one of the face coverings, again in randomised order (with the exception that the tape condition always came last). The purpose of the control condition was to obtain reference material for each talker where his/her face was not concealed by any kind of face covering. The control condition will serve as the baseline condition in the acoustic and perception experiments presented later in this thesis.

The prompting method was screen-prompting. The stimulus lists were presented in timed PowerPoint presentations on an Edge 10 H170 LCD monitor (controlled by an Acer Aspire TimelineX 3820TG notebook). One stimulus sentence (e.g. *He said [na:p].*) was presented per slide for 2.55s. Between successive sentences a black screen was shown for 0.55s. After each block of 16 stimuli a slightly longer break of 2.5s was given so as to provide the talkers with an indication of how many sentences from the current list were still to be read.

The talkers were instructed to read the stimuli carefully yet fluently, and to control their speaking style to the best of their ability. They were advised to vary as little as possible in speaking volume (loudness of their voice) and intonation (pitch contours) as they read the test sentences, and to speak clearly but not in an exaggerated manner. Speaking tempo was controlled for within the limits of the timed PowerPoint presentation. Moreover, subjects were asked to control their facial expressions to the best of their ability (neutral, no strong eyebrow raising), and to have their lips closed at the start and end of each utterance. They were furthermore advised to continue reading in case reading or pronunciation errors occurred. Misread or mispronounced stimuli were repeated at the end of each take.

3.1.5 Recording set-up

The database was recorded in a professional sound-treated TV studio at the Department of Theatre, Film and Television, University of York, United Kingdom. Participants were seated in front of a plain green background, and were asked to avoid marked head movements while the recordings were taking place. Two light sources were arranged to produce a uniform illumination across each talker’s face (the studio was windowless). They were asked to put on plain black T-shirts or hooded sweatshirts that were provided to them, and not to wear spectacles or conspicuous jewellery in order to avoid possible reflection caused by the spotlights.

As Figure 3.2 illustrates, three simultaneous continuous audio recordings were made. A DPA 4066 Omnidirectional Headband Microphone captured the speech signal approximately 2cm from the right-hand corner of each talker’s mouth. It was taped to the facewear with black or skin-coloured adhesive tape, if necessary.

Two Røde NTG-2 Dual Powered Shotgun Condenser Microphones captured the audio from 2.3m in front of and 1.4m behind the talker. The rearward microphone was placed at the height of each talker’s head and was therefore not visible in the resultant videos. Audio was recorded with an Edirol R-4 Pro Portable 4 Channel Recorder and a Sound Devices 552 Portable Production Mixer.

Two simultaneous continuous HD video recordings were made using two Panasonic AG-HPX171E Camera Recorders which were positioned so that the images consisted of the talker’s head and shoulders. The half-profile camera was placed opposite the location of the headband microphone to avoid the headset occluding part of the side of the talker’s mouth/cheeks.

The monitor for stimulus prompting was placed directly below the camera lens of the frontal camera (following Llamas *et al.*, 2008). This created the impression that the talkers were looking into the lens. The frontal camera took its audio input from the headband and the frontal microphone. The rearward microphone and the half-profile camera captured the speech signal separately. To facilitate the temporal alignment of all audio and video streams during post-processing, a clapperboard signal was given at the start of each take.

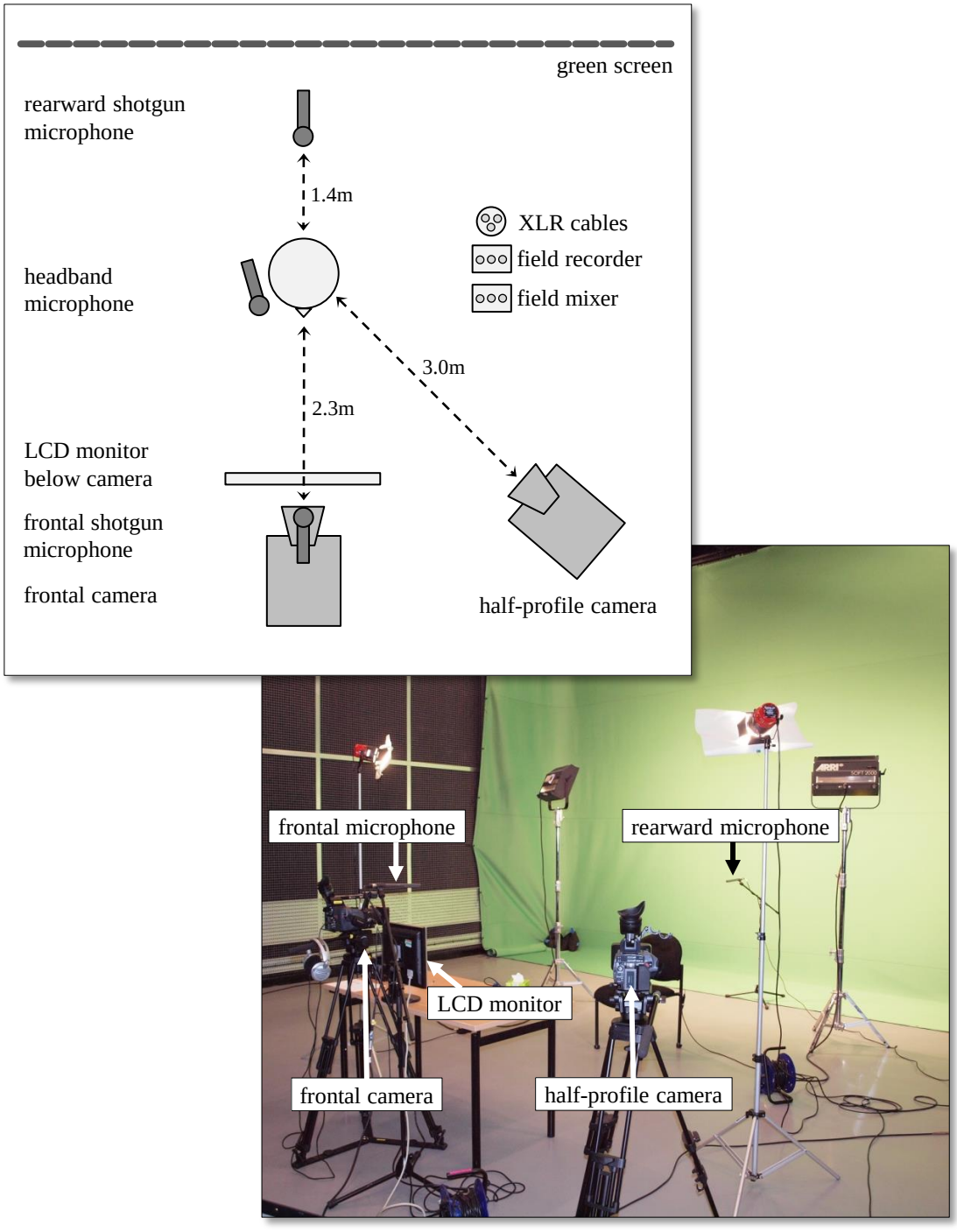

Figure 3.2. Recording set-up during data collection for the AVFC speech corpus. The audio was captured with three microphones (headband, frontal, rearward), and the video was recorded with two cameras (frontal, half-profile). The talker was seated in front of a green screen, with the face fully illuminated, and was reading the target stimuli from a computer screen placed directly below the frontal camera lens.

3.1.6 Post-processing

Several post-processing steps were necessary in order to make the collected data accessible for further experimentation. Firstly, the original multimedia container files (QuickTime File Format, approximately 419GB recorded material in total) were decoded (using MEncoder, XVID codec), and all audio and video streams were extracted and stored in separate files (audio: RIF WAV format, 48kHz sample rate, 768kbit/s bit rate, 16-bit signed integer PCM encoding; video: LAVC, XVID MPEG-4 video codec, 25 frames/second, 2024kbit/s bit rate, 24-bit sample size, 960x720 resolution). Subsequently, the relevant audio channels and video streams for each talker were identified, organised into subdirectories (for each talker and facewear conditions separately), and consistently renamed following a predefined nomenclature. A README file, which specifies the latter in more detail, is available upon request from the author of this thesis.

A subset of the collected audio data, namely the recordings of all ten talkers reading the target sentences in the control and all eight facewear conditions captured with the headband microphone, was then automatically segmented and transcribed. This required multiple pre-processing steps in *Praat 5.1.44*[15], such as the automatic labelling of pauses, editing and labelling of audio files and TextGrids, and finally the execution of the Linux-based forced aligner *MAUS*[16] (‘Munich AUtomatic Segmentation’; see e.g. Schiel, 2004). The output of this process was 5,266 separate audio files and corresponding TextGrids, consisting of one target sentence each. These data enabled, among other things, the faster extraction of suitable acoustic material for the acoustic-phonetic experiments presented in Chapter 4.

[15] Available from: http://www.goo.gl/HQQmGG [Accessed: 7th May 2014].
[16] Available from: http://www.goo.gl/XWm3qC [Accessed: 7th May 2014].

3.2 Use in this thesis

As previously noted, the primary intention behind the collection of the AVFC corpus was to provide multi-purpose auditory-visual speech data based on footage of talkers wearing facewear while producing phonetically-controlled speech material, which could subsequently be used for a range of experiments that empirically explore 'facewear effects' (as introduced in §2.1.2). The further processing of the data was guided by the specific requirements of each experiment. Additional technical specifications are given in the corresponding methodology sections in each chapter.

When designing the corpus, extra effort was put into creating a speech database which could potentially be used for future research (by the author and others) in forensic speech science and related fields of study. In recent years, several large multimodal databases have been created, such as the XM2VTSDB ('Extended Multi Modal Verification for Teleservices and Security Applications Database') or the BANCA database ('Biometric Access Control for Networked and E-Commerce Applications'). These serve the purpose of, for example, testing person recognition performance by automatic multibiometric systems (Goecke, 2005; Aleksic & Katsaggelos, 2006; Trojanová *et al.*, 2008). However, the elicited speech material is often phonetically and acoustically unsuitable for perceptual testing with human subjects, especially where the focus is on speech perception in adverse listening and/or viewing conditions. Also, and more importantly, thus far no corpus has adopted different types of facial occlusion as a within-subject design parameter (except e.g. hats in the 'Digital Audio-Visual Integrated Database'; see Mason *et al.*, 1996). For these reasons it was mandatory to collect new data for the purposes of the present research.

One major limitation of the corpus is undoubtedly the highly-controlled speech material in the form of (mainly) nonsense syllables. Future data collection of this kind should therefore include 'real' words, a larger vowel inventory, varying prosodic contexts, spontaneous speech, etc., and ideally forensically-relevant speaking styles (such as emotional, shouted or whispered speech). Also, the number of talkers is not sufficient to adequately test for talker effects, or for the performance of automatic speech/speaker recognition systems.

More positively, the number of recorded talkers in the present corpus (N = 10) is in fact higher than that used in a fair amount of other studies that look at auditory-visual speech perception. Especially studies on lip-/speechreading often use speech material elicited from only one or two talker/s (see e.g. Preminger *et al.*, 1998; Brungart & Simpson, 2005; Lidestam & Beskow, 2006; Rosenblum *et al.*, 2007). Moreover, the recordings were made in a highly controlled environment, and the resultant audio and video data are of very high quality. To increase the reusability of the data and to compensate for the relatively small set of talkers, the corpus design incorporated different microphone positions, camera angles, and the option for chroma-keying.[17]

In sum, the AVFC corpus is the first of its kind as it includes a considerable variety of face coverings. It ought to be considered a (relatively small yet high-quality) resource for further empirical studies on auditory-only, auditory-visual and visual-only speech (and face) processing. Moreover, the AVFC corpus is potentially of practical relevance to the forensic community, in that it can provide reference material for forensic phonetic and acoustic work on authentic cases involving talkers whose facial appearance is fully or partially disguised. The interested reader is invited to contact the author in order to gain access to the data.

The next chapter (Chapter 4) presents the first two experiments that made use of the collected speech data. They consist of a comparative acoustic-phonetic analysis of voiceless fricatives and plosives, which have been produced in the control (no facewear) condition and through the various face coverings listed in §3.1.2.

[17] Chroma-keying is a compositing technique for replacing a monochromatic background of a moving or still image with a different image in post-production. In the present context, chroma-keying enables the design of studies which, for example, investigate the effects of varying types of distracting (visual) backgrounds on speech processing.

4 Acoustic properties of facewear speech

4.1 Experiment 1: Voiceless fricatives

The current chapter presents the findings from an acoustic-phonetic investigation of voiceless fricatives and plosives which were produced by talkers wearing a range of face and head coverings. The data used for analysis were extracted from the AVFC corpus (see Chapter 3). The study centres on the following questions:

- Does facewear change the acoustic properties of voiceless fricatives and plosives? Specifically, are selected intensity, temporal, and spectral measures of the speech sounds modified when the fricatives and plosives have been produced while the talker's face is disguised by facewear?
- Assuming that acoustic facewear effects emerge, are the two classes of fricatives, namely the sibilants and non-sibilants, differently affected by facewear? Correspondingly, to what degree and in what manner does facewear alter the acoustic characteristics of plosives?
- Lastly, which type of face covering has, by and large, the most deleterious effect on the acoustics of fricatives and plosives?

The first part of the chapter describes the motivations for obtaining intensity and spectral measures of the four voiceless fricatives /s/, /ʃ/, /f/, and /θ/. This is followed by an overview of their most relevant articulatory and acoustic characteristics. After this, the analysis techniques and results of a thorough statistical analysis of the fricative data are presented (Fecher, 2011; Fecher & Watt, 2011).[18]

[18] Some of the results of this study were presented in 2011 at the *17th International Congress of Phonetic Sciences (ICPhS)*, the *20th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*, and the *4th ISCA Tutorial and Research Workshop on Experimental Linguistics (ExLing)*.

4.1.1 Introduction

4.1.1.1 Aim and motivation

The choice of fricatives for this study, namely /s/, /ʃ/, /f/, and /θ/, was motivated by three factors. Firstly, previous research on consonant perception has shown that fricatives are the most common source of listening errors when listeners are asked to auditorily identify a set of consonants. In Chapter 5, a range of studies of consonant recognition will be introduced (Miller & Nicely, 1955; Wang & Bilger, 1973; Redford & Diehl, 1999; Benkí, 2003; Smits *et al.*, 2003; Weber & Smits, 2003; Phatak & Allen, 2007; Woods *et al.*, 2010). Overall, these studies have established that fricatives are prone to all types of feature-processing errors, i.e., place of articulation, manner of articulation, and voicing errors. Furthermore, the labiodental and dental fricatives /f/ and /θ/ were found to be particularly difficult to identify perceptually. While the alveolar and postalveolar (palatoalveolar) fricatives /s/ and /ʃ/ are known to be well recognised even at very low signal-to-noise ratios (SNRs), /f/ and /θ/ are misidentified significantly more often, even at relatively high SNRs. The consonant recognition study to be presented in Chapter 5 is an attempt to replicate and extend the outcome of earlier studies. As will be seen, fricatives will once again bring about the highest number of misperceptions, both in the control (no facewear) condition and when they were produced through facewear. The acoustic examination of fricatives was for these perceptual reasons considered worth pursuing.

The second motivation for focusing on fricatives was the relevance of fricative analysis in forensic phonetic casework. Forensic speech scientists generally acknowledge fricatives as a valuable speaker-discriminating feature, because fricative spectra can vary greatly from talker to talker (Hayward, 2000). This was demonstrated, for example, by Haley *et al.* (2010) for spectra of English word-initial /s/ and /ʃ/. Haley and colleagues found that the spectra of different talkers can considerably overlap, to the extent that one talker's /s/ may be acoustically indistinguishable from another talker's /ʃ/ production. Similar effects were observed in recent research by Kavanagh (2013), who assessed the degree of intra- and inter-talker variability in /s/ (among other segments) produced by 30 native British English

speakers. Kavanagh examined the means and ranges of different acoustic parameters (duration, spectral moments) in different filter conditions, and found that /s/ shows great potential for discriminating between individual talkers.

Furthermore, a recently-conducted international survey of forensic speaker comparison practices by Gold & French (2011) evaluated the frequency with which individual segments (consonants and vowels) are examined by forensic practitioners. All respondents to the survey reported subjecting consonants to some form of examination, such as auditory inspection and/or acoustic analysis of timing and spectral properties. Participants provided answers in the form of 6-point Likert ratings ranging from 1 (denoting that fricatives are never analysed in casework) to 6 (always analysed). The responses, broken down by consonantal manner of articulation, revealed that fricatives were ranked highest, with a mean Likert rating of 4.85 (*SD* = 1.21; see Gold & French, 2011: 301).

Thirdly, fricatives were chosen for experimentation based on their distinctive acoustic structure. The energy distributions in higher frequency bands of the acoustic spectrum are especially discriminative for this class of sounds. Based on research by Llamas *et al.* (2008), which was presented in §2.3.1, it is hypothesised that the changes to the acoustic signal caused by facewear will be particularly prominent for fricatives. Llamas *et al.* had observed that facewear attenuated sound energy especially in higher frequency bands of the acoustic spectrum. As a result, fricatives were among the speech sounds that were most strongly affected acoustically.

In the present study, two aspects are of major interest. These are a) the impact of facewear on the acoustic properties of the two classes of fricatives (sibilants and non-sibilants), and b) the extent to which different types of facewear affect these sounds. With regard to a), it is anticipated that the two classes will be differentially affected when the fricatives are produced through facewear. This hypothesis is based on the known discrepancy between the acoustic structures of sibilants and non-sibilants. As will become apparent over the course of this chapter, the two classes can be distinguished based on their spectral shapes and energy distributions.

Considering b), it is estimated that the magnitude of acoustic modifications to fricatives caused by facewear will largely (but not exclusively) depend on the type of

facewear (material) tested. For example, the thick, sound-absorbent composite shell of the motorcycle helmet, or the double-knitted fabric of the balaclava, are expected to absorb sound energy much more heavily than the comparatively thin, lightweight textiles of the surgical mask or the *niqāb*, which are estimated to cause relatively minor spectral effects.

The scope of this work – unlike much of the preceding literature on the acoustics of fricatives – is *not* the *classification* of fricatives based on acoustic measures. Rather, the goal is to obtain an overall impression of the sound-absorbing characteristics of a variety of face-concealing garments and headgear. The acoustic effects of facewear are therefore examined for each of the four fricatives individually.

The next section provides an overview of relevant articulatory and acoustic properties of /s/, /ʃ/, /f/, and /θ/. After this, the methodology used for the acoustic measurements is described, and the results of the statistical analysis are presented.

4.1.1.2 /s ʃ f θ/ revisited

English voiceless fricatives are produced with a turbulent airstream, which is the consequence of a pulmonically-initiated egressive jet of air being channelled through a narrow constriction somewhere along the vocal tract and hitting a nearby obstacle.[19] This obstacle is the upper teeth for the production of /s/ (as in English <**s**ip>), the lower teeth for /ʃ/ (as in <**sh**e>) and /θ/ (as in <**th**in>), or the upper lip for /f/ (as in <**f**ew>). The random velocity fluctuations in the airflow, which are caused when the air is forced at high speed through the constriction, are the sound source for voiceless fricatives (Laver, 1995; Harrington, 2010; Johnson, 2003).

The acoustic consequence of the turbulent airstream is aperiodic energy produced at or near the place of maximum constriction, typically above ~1kHz and with peaks

[19] The term ‘voiceless’ is used in its phonetic sense throughout the thesis. It refers to the absence of voicing (vocal fold vibration), rather than to a phonological contrast.

above ~5kHz (Hayward, 2000; Harrington, 2010). The overall spectral shape of the emerging sounds is defined by the place, degree, and shape of the narrowest constriction (especially the length of the front cavity), and marginally by the pressure and rate of the airflow (Stevens, 1998; Harrington, 2010). The interaction between airflow and acoustic factors in fricative production is rather delicate, given that there is a critical rate of airflow through the constriction below which the airflow is laminar (non-turbulent) and relatively silent, and above which the airflow is turbulent and noisy, producing 'hissing' or 'hushing' sounds (Laver, 1995).

The search for stable acoustic correlates to distinguish between fricatives has proven challenging in the past. The set of quantitative parameters to specify the acoustic structure of fricatives still lacks standardisation somewhat. There is as yet no uniquely-defined list of parameters by which to characterise articulatory and perceptually-relevant acoustic cues to fricatives (Flipsen *et al.*, 1999; Jongman *et al.*, 2000). For example, no single metric has been found to reliably classify the place of articulation (Tomiak, 1990; Jongman *et al.*, 2000; Munson, 2001; Blacklock & Shadle, 2003), or to distinguish between fricatives in a talker-invariant manner (Haley *et al.*, 2010).

Nonetheless, there now exists a large set of acoustic parameters which can be used to distinguish between fricatives quite effectively. These can be grouped as intensity, temporal, and spectral measures. Regarding the last of these, researchers frequently examine formant transitions, the location of the spectral peak, and gross spectral shapes (Jongman *et al.*, 2000; Tabain & Watson, 1996; Maniwa *et al.*, 2009; Haley *et al.*, 2010).

The spectral shape of fricatives is typically parameterised by spectral moments. Spectral moment analysis is a statistical procedure in which local (mean frequency, standard deviation) and global (skewness, kurtosis) aspects of the spectral distribution of a sound are captured. The method dates back to Forrest *et al.* (1988), who calculated a series of FFT (Fast Fourier Transform) spectra from the onsets of word-initial voiceless obstruents (fricatives, plosives, affricates). They then treated each FFT spectrum as a random probability distribution from which the spectral

moments were derived. The first four spectral moments will be introduced in further detail over the course of this chapter.

As noted before, fricatives can be classified as 'sibilants' or 'non-sibilants'. The two classes are characterised by different acoustic patterns *between* the classes, but similar acoustic properties *within* each class. Figures 4.1 and 4.2 illustrate that sibilants and non-sibilants differ greatly in terms of the intensity of the frication noise. Sibilants are specified by substantially (i.e., ~10–15dB) greater intensity than non-sibilants (see the darker shadings in Figure 4.1, and higher mean sound pressure levels in Figure 4.2). This is the result of the larger front cavity for the production of sibilants, and the airstream hitting the teeth and producing high-energy turbulence. Within each class of fricatives, the overall intensities are not considerably different from each other. Interestingly, reducing the amplitude of sibilants can lead them to be perceived as non-sibilants, but not vice versa (Harrington, 2010). This adumbrates the fact that the two classes not only differ in terms of their overall intensities, but can be discriminated on the basis of the noise spectrum.

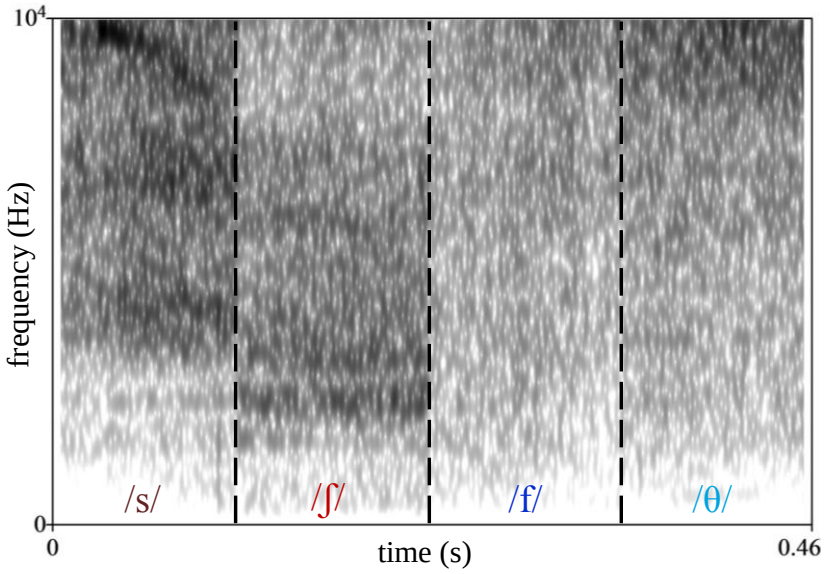

Figure 4.1. Wideband spectrogram showing, from left to right, steady-state phases of the sibilants /s/ and /ʃ/, and the non-sibilants /f/ and /θ/, each spoken in syllable onset position (before /ɑː/) by one of the male talkers recorded for the AVFC corpus.

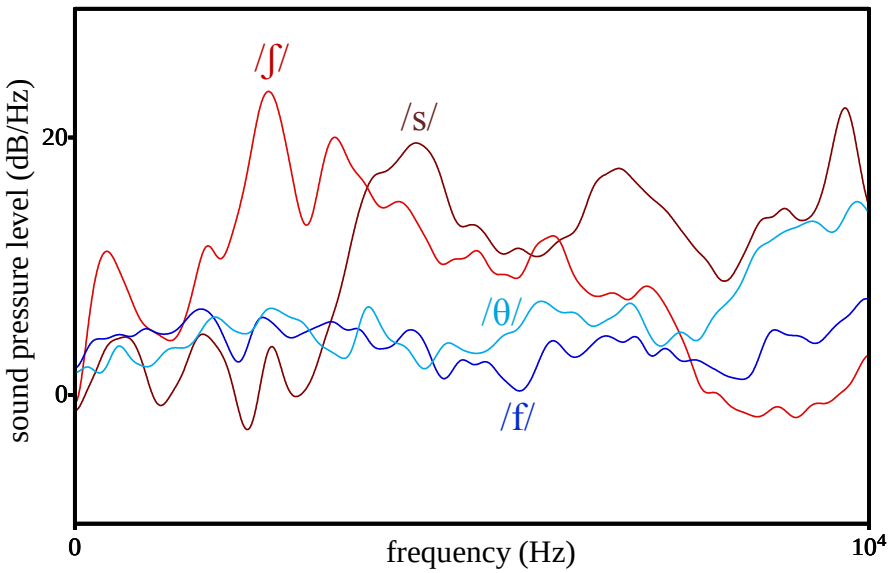

Figure 4.2. Cepstrally-smoothed power spectra for the sibilants /s/ and /ʃ/, and the non-sibilants /f/ and /θ/, each spoken in syllable onset position (before /ɑː/) by one of the male talkers recorded for the AVFC corpus (dark red = /s/, light red = /ʃ/, dark blue = /f/, light blue = /θ/).

Considering the sibilants more closely, it can be noted that compared to /ʃ/, the front cavity is shorter, and the constriction is further away from the lips, during the production of /s/. Furthermore, the air hits the upper teeth when /s/ is produced, which creates high-energy turbulence, especially in higher frequency bands. The sibilant /ʃ/, on the other hand, is typically accompanied by lip-rounding and a larger sublingual cavity (which effectively lengthens the front tube). Acoustically, these articulatory differences result in the sound source for /s/ being filtered by front cavity resonances, with the consequence that the peak and mean frequencies are on average 2–4kHz higher for /s/ than for /ʃ/ (Johnson, 2003; Onaka & Watson, 2000; Harrington, 2010). Much of the spectral energy for /s/ is concentrated in the F4 or F5 range, or higher (3.5–5kHz, 6–8kHz), while /ʃ/ gives rise to prominent mid-frequency peaks in the F3 and F4 ranges (Stevens, 1998; Hayward, 2000; Jongman *et al.*, 2000; Johnson, 2003; Harrington, 2010; Stevens, 2010). Moreover, the two sibilants differ from each other with respect to the overall slope of their spectra. As Figure 4.2 shows, the curve for /ʃ/ rises steeply to its peak, while it rises more gradually for /s/.

For the generation of non-sibilants, the length of the front cavity is negligible. The overall intensity is usually low (see Figure 4.1). The spectra of non-sibilants are often described as ‘flat’ or ‘diffuse’ (see Figure 4.2). This implies that the acoustic energy (below ~10kHz) is distributed throughout the spectrum without major resonances (peaks) or regions of prominence (Johnson, 2003; Harrington, 2010). Due to their greater spectral diffuseness (large variance), spectral moments cannot reliably differentiate between /f/ and /θ/ (Forrest *et al.*, 1988; Shadle & Mair, 1996; Jongman *et al.*, 2000; Harrington, 2010).

Following the presentation of the aims and motivations for the study, and the brief overview of the articulatory and acoustic properties of fricatives, the next sections report on the applied methodology.

4.1.2 Method

4.1.2.1 Talkers and facewear

From the ten native British English speakers recorded for the AVFC corpus (see Chapter 3), three females and three males were selected at random. However, the data for one of the male talkers were excluded prior to this selection due to his atypical pronunciation of /ʃ/ (possibly laminal articulation with tongue tip raised, lack of tongue blade grooving and lip rounding). The mean age of the six selected talkers was 25.7 years (*SD* = 6.1). To recall, all of them were phonetically trained and familiar with the IPA, which enabled them to produce the target stimuli reliably and consistently (within the bounds of what is feasible). Including different talkers in this study compensated (at least to some extent) for inter-talker variability (Haley *et al.*, 2010; Kavanagh, 2013). Auditory inspection of the selected acoustic material confirmed that none of the talkers produced the fricatives of interest in any (for British English) unconventional or non-standard way.

Of the eight types of facewear included in the AVFC corpus, seven were chosen for testing. The balaclava with the mouth hole was excluded from this study, because it was expected that it would have no (or only very little) impact on the acoustics of fricatives. Given that the mouth is not occluded for this type of mask, the oral airstream is fully maintained during speech production (unlike for all other tested face coverings).[20]

4.1.2.2 Speech material

The speech material consists of the sibilants /s/ and /ʃ/, and the non-sibilants /f/ and /θ/. These were extracted from the /C_1ɑːC_2/ syllables (embedded phrase-finally in the carrier phrase *He said [stimulus]*) recorded for the AVFC corpus. For each fricative, two tokens were selected per syllable position, i.e., two per onset (/C_1/) and two per coda (/C_2/). The samples for each talker were chosen at random from the list of automatically segmented and forced-aligned headband microphone recordings (as specified in §3.1.6). Auditory inspection of the material ensured that no mispronunciations or reading errors occurred in the preselected files. Where this was the case, the corresponding samples were excluded and replaced.

The acoustic measurements were performed on the audio recordings captured with the DPA 4066 Omnidirectional Headband Microphone (48kHz, 768kbit/s, 16-bit signed integer PCM encoding). As the reader will recall (see §3.1.5), the microphone was placed approximately 2cm from the right-hand corner of the talker's mouth, and taped to the outer surface of the mask, where necessary.

In total, 768 fricative samples were selected and hand-segmented: 6 talkers x 4 fricatives x 2 syllable positions x 2 tokens x 1+7 facewear conditions (control + 7 types of facewear).

[20] The nose, on the other hand, is fully covered by the mask. This type of balaclava would therefore be worthwhile examining with regard to its effect on nasals and nasalised sounds. Some remarks on the impact of facewear on the perception of nasals are given in §6.2.3.1.

4.1.2.3 Procedure

The fricative portions were manually segmented and automatically extracted from the /C_1ɑ:C_2/ syllables by means of a *Praat* script in *Praat 5.1.44*. The segmentation points were based on auditory and visual inspection following established procedures described in the literature (Shadle & Mair, 1996; Jongman *et al.*, 2000; Machač & Skarnitzl, 2009; Haley *et al.*, 2010). The analysis was based on the steady-state phase of each fricative, which is the section where the articulation of the segment was momentarily held (Laver, 1995). The fricative samples were on average 253ms long (*SD* = 73ms).

Measurements were taken (in *Praat*) from wideband spectrograms (Gaussian; window length = 5ms), and more specifically, from averaged FFT spectra rather than from spectral slices (bandwidth = 500Hz; cepstral smoothing applied). The latter decision was based on work by Tabain & Watson (1996), who report considerable differences between these two types of analysis. In their data, averaged FFT spectra yielded better results, e.g. for the classification of /θ/. By contrast, Gordon *et al.* (2002) found that using spectral slices can reduce coarticulation effects. In the present data, however, coarticulation was not considered an issue, because the nucleus in the tested CVC syllables was consistently the open back vowel /ɑ:/, and the carrier sentence preceding /C_1/ was always *He said*.

No pre-emphasis filter was implemented in the present study (following e.g. Tabain & Watson, 1996), and the averaged FFT power spectra were computed over non-filtered speech. The audio was sampled at 48kHz, which allowed the frequency range up to 24kHz to be taken into account (Forrest *et al.*, 1988; Jongman *et al.*, 2000; Gordon *et al.*, 2002; Jones & Llamas, 2008; Haley *et al.*, 2010; Kavanagh, 2013). This was considered beneficial because fricatives have been shown to encode place information above the classically employed 10kHz cut-off point (Shadle & Mair, 1996; Tabain & Watson, 1996; Hayward, 2000).[21]

[21] The results of studies which take frequencies >10kHz into account vary. For example, Tabain & Watson (1996) found that useful acoustic information, e.g. for /f/, may in fact be encoded in the spectrum above 12kHz. Tabain (1998), however, points out that acoustic cues

In sum, for each of /s/, /ʃ/, /f/, and /θ/, the following spectral properties and intensity of the frication noise were measured:

A. Intensity measure
    - mean intensity (in decibels)
B. Spectral measures
    - spectral peak (in Hertz)
    - centre of gravity (in Hertz)
    - standard deviation (in Hertz)
    - skewness (dimensionless)
    - kurtosis (dimensionless)

4.1.3 Results

The statistical analysis of the data was carried out by means of a series of three-way repeated-measures analyses of variance (ANOVAs) using *IBM SPSS Statistics V.19.0.0.1*.[22] The dependent factors under consideration were 'intensity', 'spectral peak', and the first four statistical moments of the FFT spectra, namely 'centre of gravity' (CG), 'standard deviation' (SD), 'skewness', and 'kurtosis'. The independent within-subject factors were 'fricative' (/s/, /ʃ/, /f/, /θ/), 'facewear' (control, balaclava without mouth hole, helmet, hoodie/scarf, *niqāb*, rubber mask, surgical mask, tape), and 'syllable position' (onset, coda). There were also two between-subject factors, namely 'talker' and 'gender'. The results are reported in the form of averages across the speech data elicited from all six talkers. Gender effects are recaptured in §4.1.3.1. Note that in the illustrations shown in this chapter, the

above 10kHz cannot reliably distinguish between /f/ and /θ/. On the other hand, Shadle & Mair (1996) analysed /s/, /ʃ/, /f/, and /θ/ spectra with frequency ranges up to 17kHz, and observed that spectral moments were considerably affected by the frequency range used. They suggest that filtering causes an artificial cut-off in the spectral content, which potentially distorts spectral measures.

[22] Available from: http://www.goo.gl/3b8L0A [Accessed: 7th May 2014].

balaclava (without the mouth hole) will appear as ‘balaclava 1’ so as to be consistent with the naming conventions in later chapters.

Effects are reported as significant when $p < .05$. Where Mauchly’s test indicated that the assumption of sphericity had been violated, the degrees of freedom, *p*-values and effect sizes (η_p^2) were adjusted using the Greenhouse-Geisser correction (the correction factor ε is listed in the corresponding results table in such cases).

4.1.3.1 Overview

To begin with, the effects of fricative, facewear and syllable position were analysed for each dependent variable separately. The results of this analysis are shown in Appendix D.5 (see Table D.39).

As expected, there was a highly significant main effect of fricative on all dependent variables (*p*s < .001). The effect of facewear was significant for intensity, CG (*p*s < .001), skewness, and kurtosis (*p*s < .05), but not for spectral peak ($p = .500$) and SD ($p = .583$). However, given that the interaction between fricative and facewear was significant for spectral peak ($p < .05$) and SD ($p < .01$), it was decided to pursue the statistical analysis of all four fricatives.

There was, moreover, a significant effect of syllable position on intensity [$F(1,5) = 47.96$, $p < .01$, $\eta_p^2 = .91$], spectral peak [$F(1,5) = 38.45$, $p < .01$, $\eta_p^2 = .89$], CG [$F(1,5) = 122.65$, $p < .001$, $\eta_p^2 = .96$], and skewness [$F(1,5) = 15.25$, $p < .05$, $\eta_p^2 = .75$]. No overall significance was found for SD and kurtosis. However, ANOVAs performed per facewear condition and fricative revealed that the effect of syllable position was significant in some of the facewear conditions for SD and kurtosis.

Owing to the overall significant effect of syllable position, it was decided to divide the dataset into the fricatives produced in onset position, and those extracted from the coda position of the target stimuli. For want of space, only the results for the *onset* data are reported in this thesis. This decision was guided by the speech perception

studies that will be presented in Chapters 5 and 6, which exclusively take onset consonants into consideration.

Finally, statistical analysis showed no evidence of a gender effect for most dependent variables, i.e., intensity (p = .755), spectral peak (p = .214), standard deviation (p = .824), skewness (p = .083), and kurtosis (p = .562). However, a significant (and expected) gender effect (p < .001) was observed for centre of gravity. The fricatives produced by the females on average had higher centres of gravity (see Tufts & Frank, 2003; Llamas *et al.*, 2008; Pepiot, 2012). The effect was significant in the control, helmet (ps < .001), hoodie/scarf, *niqāb*, surgical mask, and tape conditions (ps < .01), but not in the balaclava (p = .421) and rubber mask (p = .074) conditions.[23]

After the dataset was split up by syllable position, and the statistical analysis was repeated (see Appendix D.5, Table D.40), a significant main effect of (onset) fricative on intensity [$F(3,18) = 1824.60$, $p < .001$, $\eta_p^2 = .99$], spectral peak [$F(3,18) = 37.07$, $p < .001$, $\eta_p^2 = .86$], CG [$F(3,18) = 472.48$, $p < .001$, $\eta_p^2 = .99$], SD [$F(3,18) = 626.51$, $p < .001$, $\eta_p^2 = .99$], skewness [$F(3,18) = 242.87$, $p < .001$, $\eta_p^2 = .98$], and kurtosis [$F(2,11) = 82.32$, $p < .001$, $\eta_p^2 = .93$] was found. This implies that averaged across facewear conditions, the four fricatives significantly differed from each other with respect to all acoustic properties measured.

There was, moreover, a significant main effect of facewear on intensity [$F(7,42) = 346.62$, $p < .001$, $\eta_p^2 = .98$], spectral peak [$F(7,42) = 7.10$, $p < .001$, $\eta_p^2 = .54$], CG [$F(7,42) = 25.94$, $p < .001$, $\eta_p^2 = .81$], and skewness [$F(7,42) = 4.01$, $p < .01$, $\eta_p^2 = .40$], but not on SD (p = .412) and kurtosis (p = .153). This suggests that, averaged across fricatives, the different types of facewear significantly affected most acoustic properties of the fricatives.

The interaction between fricative and facewear was significant for intensity [$F(21,126) = 6.73$, $p < .001$, $\eta_p^2 = .53$], spectral peak [$F(21,126) = 7.95$, $p < .001$, η_p^2

[23] The author acknowledges that gender is an important factor in phonetic research, and that the acoustic analysis of female and male speech ought to be carried out separately. However, as a gender effect only emerged at some levels of comparison, and because a further division of the dataset would have resulted in small sample sizes, it was decided to report the results averaged across all talkers. Despite this potential limitation, the author is confident that the study can provide valuable first insights into the acoustic effects of facewear on speech.

= .57], CG [$F(21,126) = 8.63$, $p < .001$, $\eta_p^2 = .59$], SD [$F(21,126) = 3.53$, $p < .001$, $\eta_p^2 = .37$], skewness [$F(21,126) = 6.62$, $p < .001$, $\eta_p^2 = .53$], and kurtosis [$F(21,126) = 1.68$, $p < .05$, $= \eta_p^2 = .22$]. This indicates that the extent to which a particular mask modified a certain acoustic-phonetic property was dependent upon the fricative tested. In other words, different fricatives were differently affected by facewear.

To explore the significant interactions further, the data were subsequently analysed for each type of fricative and facewear (+ control) individually. In the following sections, the results of this analysis will be presented for each dependent variable separately. This is in accordance with the goals of this study, which is to explore the extent to which a particular fricative measure obtained in the control condition differs from the corresponding value in each of the facewear conditions.

4.1.3.2 Intensity

The intensity of the fricatives spoken in all facewear conditions was obtained in *Praat* using the 'To Intensity' function (minimum pitch = 70Hz; time step = 0s; mean pressure subtracted to take account of DC offset), and then computing the mean (in dB) of the intensity values of the frames within the entire segment (unit of averaging method = sones). The result of these calculations is shown in Figure 4.3.

The statistical analysis of the intensity data revealed that facewear, on average, significantly affected the intensity of all fricatives. The main effect of facewear on intensity was significant for /s/ [$F(7,42) = 96.51$, $p < .001$, $\eta_p^2 = .94$], /ʃ/ [$F(7,42) = 93.90$, $p < .001$, $\eta_p^2 = .94$], /f/ [$F(7,42) = 63.91$, $p < .001$, $\eta_p^2 = .91$], and /θ/ [$F(7,42) = 122.31$, $p < .001$, $\eta_p^2 = .95$].

As expected, the intensity of the sibilants (across facewear conditions) was on average approximately 10dB higher than the intensity of the non-sibilants. By and large, the intensities of the sibilants /s/ and /ʃ/ were more similar to each other (higher for /s/ than /ʃ/) than the intensities of the non-sibilants /f/ and /θ/ were to each other (higher for /θ/ than /f/). The relatively parallel lines in Figure 4.3 suggest that

the facewear-induced intensity changes to the fricatives were fairly consistent across the various facewear conditions.

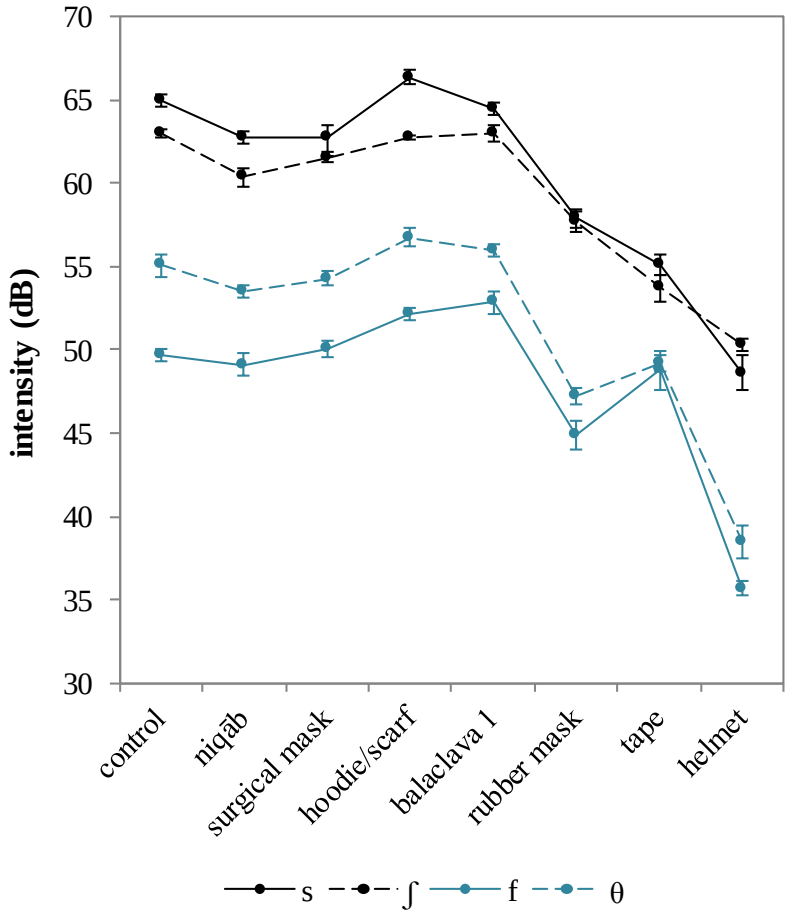

Figure 4.3. Intensity (in dB) of /s/, /ʃ/, /f/, and /θ/ produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. Note that the values on the y-axis start at 30dB instead of zero. The error bars show the standard error of the mean.

Figure 4.3 further reveals that the *niqāb*, surgical mask, hoodie/scarf, and balaclava did not provoke large intensity changes relative to the control condition. The weak acoustic effect was expected for the former two face coverings, given that they consisted of rather lightweight, low sound-absorbing materials. The results for the balaclava and hoodie/scarf, on the other hand, were less predictable. Both types of masks were manufactured from heavier, sound-absorbing fabrics. Interestingly, these two masks nevertheless caused a slight *increase* in intensity. This amplification of the frication noise may have been the result of the talkers actively compensating for

having their mouths covered up by speaking more loudly (see §4.3.2 for further discussion). This finding contrasts with Llamas *et al.* (2008), who found that the intensity of the frication noise of /s/ was lower when /s/ was spoken through a balaclava made of knitted acrylic (which may explain that /s/ was misperceived as /f/).

To examine whether the measures per fricative in the facewear conditions significantly differed from those obtained in the control condition, respectively, *post-hoc* Bonferroni-adjusted pairwise comparisons were carried out. The intensity values of the control samples for each fricative were contrasted with the corresponding values in each of the seven facewear conditions. These tests revealed that the intensity of /f/ spoken in the hoodie/scarf and balaclava conditions was significantly higher than the intensity of /f/ spoken in the control condition (ps < .01).

The effect of facewear on intensity was most noticeable in case of the rubber mask, helmet, and tape. *Post-hoc* comparisons revealed that intensity significantly dropped when /s/, /ʃ/, and /θ/ were spoken through these three types of facewear (ps < .001), and when /f/ was produced through the helmet (p < .001) or rubber mask (p < .01). All other levels of comparison produced effects that were not significant (possibly due to the comparatively small sample size).

4.1.3.3 Spectral peak

The spectral peak is the local energy maximum of the spectrum. Fricatives can have several peaks (see Figure 4.2). From utterance to utterance, one or the other of these peaks may have the greatest amplitude (Johnson, 2003). The identification of the spectral peak is therefore not as straightforward as it may appear at first. The peaks in the present study were obtained in *Praat* using the 'To Ltas (1-to-1)' function, and then calculating the maximum frequency (in Hz) associated with the maximum energy density (interpolation method = cubic). The resultant peak values were manually corrected in order to remove extreme outliers. Figure 4.4 shows the results of this procedure.

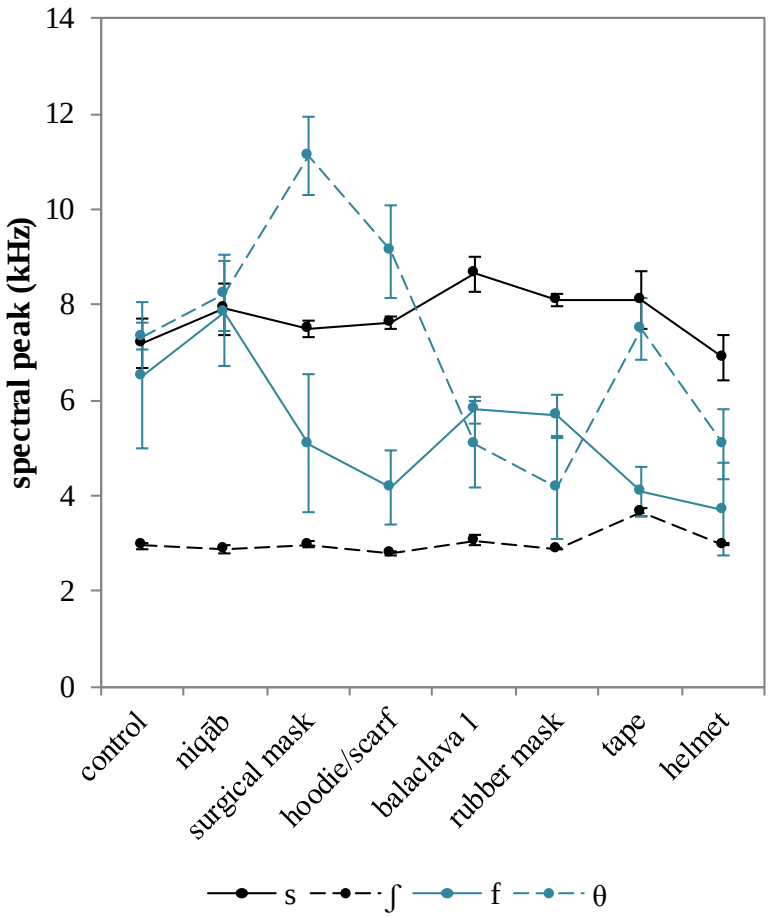

Figure 4.4. Spectral peak (in kHz) of /s/, /ʃ/, /f/, and /θ/ produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. The error bars show the standard error of the mean.

The statistical analysis revealed a significant main effect of facewear on the spectral peak of all fricatives, namely /s/ [$F(7,42) = 3.25$, $p < .01$, $\eta_p^2 = .35$], /ʃ/ [$F(7,42) = 31.86$, $p < .001$, $\eta_p^2 = .84$], /f/ [$F(7,42) = 3.14$, $p < .01$, $\eta_p^2 = .34$], and /θ/ [$F(7,42) = 18.59$, $p < .001$, $\eta_p^2 = .76$].

Altogether, the sibilants were much less affected by the acoustic facewear effects than the non-sibilants. As anticipated on the basis of the literature, the sibilants were characterised by more clearly-defined peaks than the non-sibilants. There was less variation across the sibilant samples (see the small error bars in Figure 4.4). Furthermore, the /s/ and /ʃ/ spectra varied in terms of the frequency location of the peak. This is in line with previous studies, which suggest that as the place of articulation of fricatives moves from front to back, the energy peaks shift from

higher to lower frequencies (Johnson, 2003). Here, /s/ peaked at ~7–9kHz, while the peak of /ʃ/ was much lower, at ~3kHz.

More importantly, the peak measures for /ʃ/ were similar across facewear conditions, with an increase of the peak frequency only emerging in the tape condition. *Post-hoc* Bonferroni-adjusted pairwise comparisons revealed that this difference between control and tape is significant ($p < .001$). The results for /s/ were slightly more variable than those for /ʃ/ (+ higher error bars). However, the spectral peak was overall only marginally affected when /s/ was spoken through either of the face masks. Statistically, only the peak of /s/ produced in the *niqāb* condition significantly differed from the control measures for /s/ ($p < .01$).

The peak frequencies of the non-sibilants were more difficult to determine owing to their flat, diffuse spectra. The spectral diffuseness explains both the large error bars visible in Figure 4.4 (indicative of considerable variation across samples), and the variable patterns for /θ/ and /f/ across facewear conditions. The high standard errors may indicate high inter-talker variability in the data (see e.g. Jongman *et al.*, 2000, who report that the location of the spectral peak can be talker-dependent). By trend, the peak was higher for /θ/ than for /f/. *Post-hoc* comparisons were significant only when /θ/ in the baseline was compared to /θ/ spoken through the surgical mask ($p < .01$). Altogether, the highly variable patterns make it difficult to derive any clear trends from the non-sibilant data.

4.1.3.4 Centre of gravity

The difficulty of reliably determining the spectral peak has led to the development of 'centre of gravity' techniques for the characterisation of fricatives. The centre of gravity (CG) is the first spectral moment of the spectral distribution. It expresses the frequency at which the spectral energy is predominantly concentrated, and is thus related but not equal to the peak. The CG is the point at which the energy under the curve on either side is equal (Jongman *et al.*, 2000; Johnson, 2003; Stuart-Smith *et*

al., 2003; Harrington, 2010). Here, the CG (in Hz) was measured in *Praat* by means of the 'Get centre of gravity' function (power = 2). The outcome of this procedure is shown in Figure 4.5.

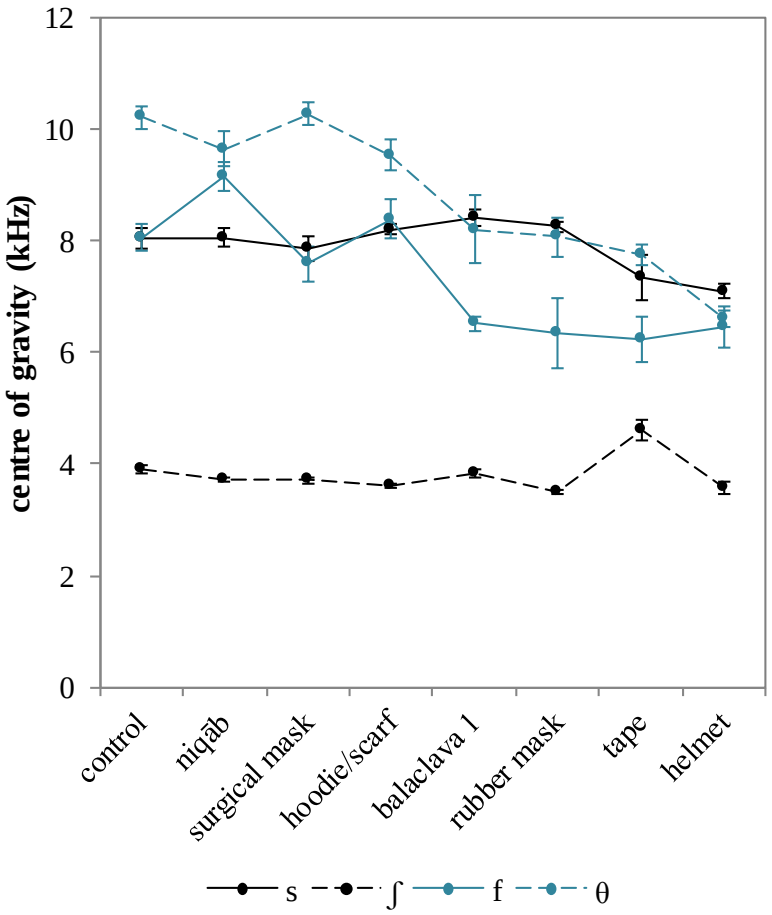

Figure 4.5. Centre of gravity (in kHz) of /s/, /ʃ/, /f/, and /θ/ produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. The error bars show the standard error of the mean.

The statistical analysis showed that there was a significant main effect of facewear on the CG of /s/ [$F(7,42) = 7.01$, $p < .001$, $\eta_p^2 = .54$], /ʃ/ [$F(7,42) = 16.28$, $p < .001$, $\eta_p^2 = .73$], /f/ [$F(7,42) = 9.75$, $p < .001$, $\eta_p^2 = .62$], and /θ/ [$F(7,42) = 17.54$, $p < .001$, $\eta_p^2 = .75$].

As can be foreseen from the formal relationship between the CG and spectral peak, the results generally mirror those reported for the peak in the preceding section. The centre frequency was higher for /s/ (at ~8kHz) than for /ʃ/ (at ~4kHz), which is again

predictable based on articulatory-to-acoustic mapping (Johnson, 2003; Harrington, 2010). As the fairly horizontal lines and small error bars for the sibilants in Figure 4.5 illustrate, the CG was only minimally affected when the speech was produced through facewear. Pairwise comparisons revealed that only /s/ produced through the helmet significantly differed from the /s/ control samples ($p < .01$). For /ʃ/, comparisons were significant when the control was compared to the tape ($p < .01$).

Regarding the non-sibilants, the CG was consistently higher for /θ/ (at ~7–10kHz) than for /f/ (at ~6–9kHz). With very few exceptions, the CG decreased when the non-sibilants were produced behind a face covering. This implies that sound energy was absorbed in particular in higher frequency bands of the spectrum (see §4.3.3 for further discussion). This effect was again most prominent in the balaclava, hoodie/scarf, helmet, and tape conditions.

Compared to the sibilants, the error bars for the non-sibilants were again higher, and the CG measures generally more variable. However, on the whole there was considerably less variation than in the spectral peak measures. For /f/, *post-hoc* tests were significant when the control samples were compared to /f/ produced through the balaclava ($p < .01$). No significant effect was found for the hoodie/scarf, helmet, and tape conditions, possibly due to the high standard errors. For /θ/, comparisons were significant only when the control samples were contrasted with the helmet, tape ($ps < .001$), and rubber mask samples ($p < .01$).

4.1.3.5 Standard deviation

The second spectral moment, namely the variance of the spectral distribution, and its positive square root, the standard deviation, is a measure of how distributed the energy is along the frequency axis. In other words, the standard deviation (SD) specifies the bandwidth of energy on either side of the mean (Jongman *et al.*, 2000; Stuart-Smith *et al.*, 2003; Harrington, 2010). Here, the SD (in Hz) was computed

with the 'Get standard deviation' function (power = 2) in *Praat*. The results are plotted in Figure 4.6.

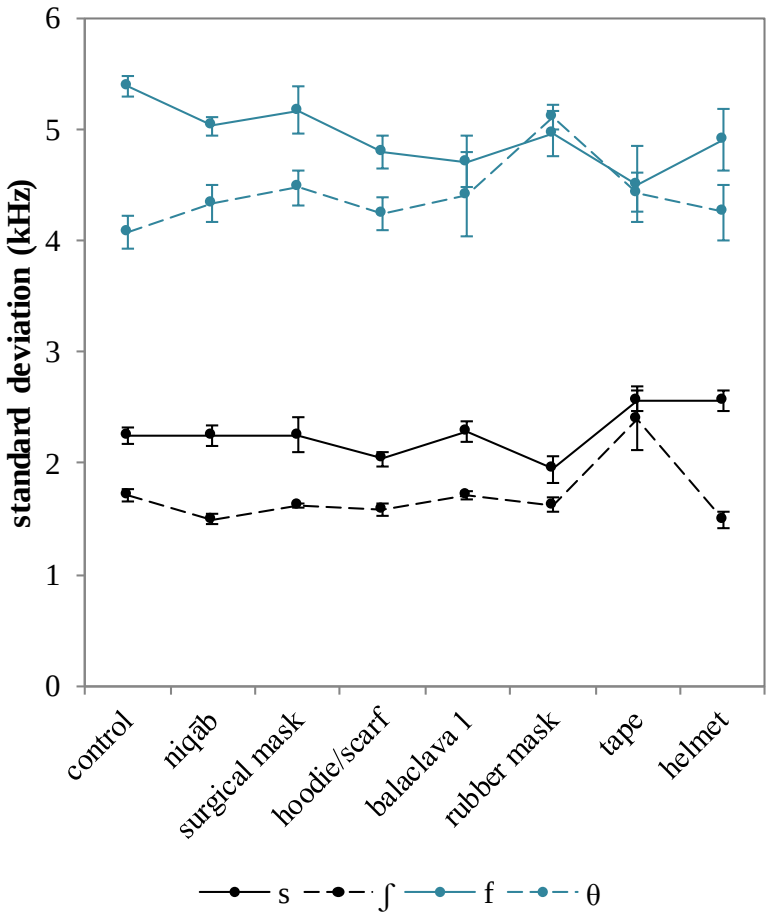

Figure 4.6. Standard deviation (in kHz) of /s/, /ʃ/, /f/, and /θ/ produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. The error bars show the standard error of the mean.

The main effect of facewear on SD was significant for /s/ [$F(7,42) = 4.11$, $p < .01$, $\eta_p^2 = .41$], /ʃ/ [$F(7,42) = 7.08$, $p < .001$, $\eta_p^2 = .54$], and /θ/ [$F(7,42) = 2.26$, $p < .05$, $\eta_p^2 = .27$], but not for /f/ ($p = .143$).

The SD of the sibilants (~1.5–2.5kHz) was considerably lower than the SD of the non-sibilants (~4–5.5kHz). The error bars in Figure 4.6 are also slightly smaller for the sibilants. These differences can be predicted from the spectral shapes typical for these sounds (see Figure 4.2). Sibilant spectra tend to be more compact, with energy

concentrated around a particular frequency, and non-sibilant spectra are relatively more diffuse (Shadle & Mair, 1996; Harrington, 2010).

Overall, the influence of facewear on the SD measures was only marginal, with the exception of the tape and helmet conditions for /s/, the tape condition for /ʃ/, and the rubber mask condition for /θ/. *Post-hoc* tests revealed that /ʃ/ spoken in the control condition significantly differed from /ʃ/ produced through the tape ($p < .01$), and that the baseline /θ/ significantly differed only from /θ/ spoken through the rubber mask ($p < .001$) or surgical mask ($p < .01$).

4.1.3.6 Skewness

Skewness is the third spectral moment of the fricative spectra. It is an indicator of the (a)symmetry (overall slant) of the energy distribution (see Figure 4.7).

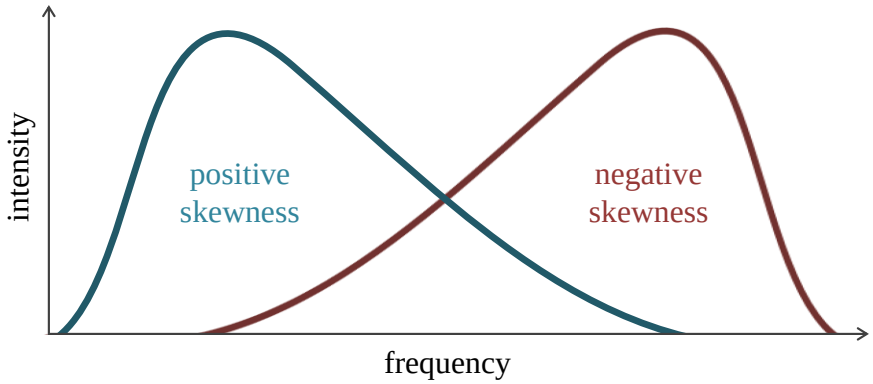

Figure 4.7. Illustration of skewness, an indicator of the asymmetry of a distribution relative to a Gaussian distribution (where skewness = 0). A spectral distribution is positively skewed when the acoustic energy is concentrated in low frequencies (negative spectral tilt), and negatively skewed when the energy is accumulated in high frequencies (positive spectral tilt).

Skewness is correlated with the CG in that it expresses how much the shape of the distribution below and above the CG differs (Jongman *et al.*, 2000; Stuart-Smith *et al.*, 2003; Harrington, 2010). Skewness values (dimensionless) are positive when the

energy is primarily concentrated in low frequency bands (negative spectral tilt), and negative when the energy is predominantly found in higher frequencies (positive spectral tilt). A value of zero denotes a normal (Gaussian) distribution, i.e., no difference in energy around the CG (Harrington, 2010). Here, measurements were taken in *Praat* using the 'Get skewness' function (power = 2). The result of these calculations is shown in Figure 4.8.

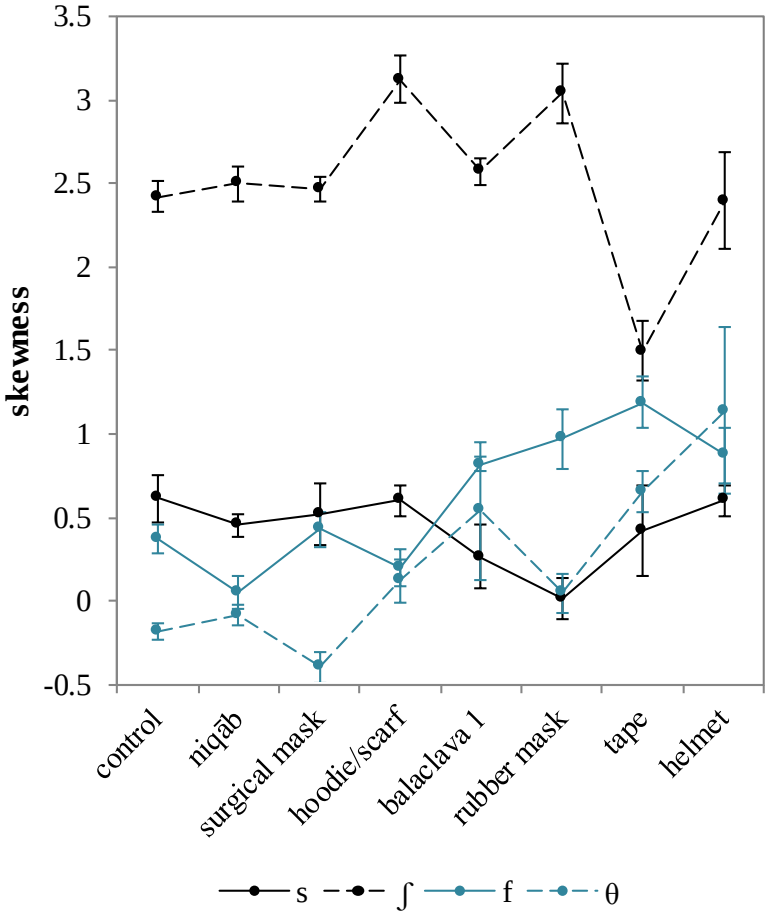

Figure 4.8. Skewness (dimensionless) of /s/, /ʃ/, /f/, and /θ/ produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. Note that the values on the y-axis start at –0.5 instead of zero. The error bars show the standard error of the mean.

Once again, statistical evaluation showed that there was a significant main effect of facewear on skewness of all fricatives, i.e., /s/ [$F(7,42) = 2.24$, $p < .05$, $\eta_p^2 = .27$], /ʃ/ [$F(7,42) = 9.58$, $p < .001$, $\eta_p^2 = .62$], /f/ [$F(7,42) = 10.71$, $p < .001$, $\eta_p^2 = .64$], and /θ/ [$F(7,42) = 4.35$, $p < .001$, $\eta_p^2 = .42$].

As expected, the highest skewness values (>2) of all fricatives were measured for /ʃ/ (except tape). This conforms to the low CG measures previously reported for /ʃ/, and indicates that the acoustic energy was predominantly skewed to the left of the spectrum (i.e., towards lower frequencies). Skewness of /ʃ/ changed markedly from the control to the hoodie/scarf and rubber mask conditions (both higher values), and from control to tape (lower values). *Post-hoc* tests revealed that skewness of /ʃ/ in the control condition significantly differed from skewness of /ʃ/ produced through the rubber mask and tape (*p*s < .01).

Accordingly, skewness for /s/ was considerably lower than for /ʃ/ (<1). The patterns across facewear conditions were more uniform for /s/, except for the marked drop in skewness from the control to the balaclava, rubber mask, and tape conditions (values approximating zero, i.e., a normal distribution of energy across the frequency range). This effect was statistically significant in case of the rubber mask ($p < .01$).

Skewness values for the non-sibilants were very low in the control condition (positive but <0.5 for /f/, negative for /θ/), indicating that a large amount of acoustic energy was concentrated in high frequency regions (see high CGs reported in §4.1.3.4). The patterns for the non-sibilants were again more variable across facewear conditions. In general, the spectrum was skewed more positively across facewear conditions. This means that compared to the control condition, relatively more energy was accumulated in lower frequency bands when facewear was involved. This effect was most evident for the balaclava, rubber mask, tape, and helmet in case of /f/. *Post-hoc* comparisons showed that these differences were significant in the helmet and tape conditions (*p*s < .01). Moreover, skewness significantly dropped in the *niqāb* condition ($p < .01$). For /θ/, the spectral distribution was skewed towards lower frequencies in the balaclava, tape, and helmet conditions. All statistical comparisons were significant for the tape ($p < .001$).

4.1.3.7 Kurtosis

The fourth spectral moment of the fricative spectra is kurtosis. Like skewness, kurtosis specifies the shape of the spectral distribution (see Figure 4.9).

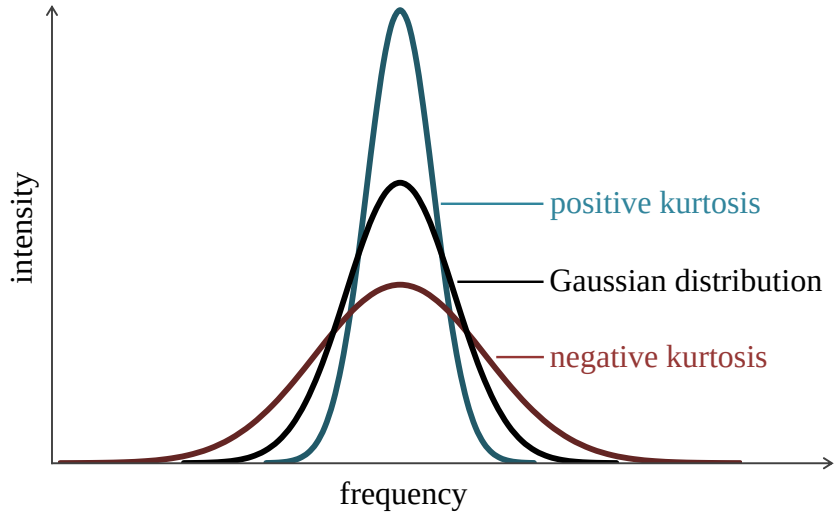

Figure 4.9. Illustration of kurtosis, an indicator of the 'peakedness' of a distribution relative to a Gaussian distribution (where kurtosis = 0). Kurtosis is positive for highly peaked distributions, and negative for relatively flat distributions.

Specifically, kurtosis is an indicator of the 'peakedness' of the distribution, i.e., it expresses to what extent the spectral energy is concentrated in a peak relative to low and high frequencies (Jongman *et al.*, 2000; Stuart-Smith *et al.*, 2003; Harrington, 2010). Kurtosis values (dimensionless) are positive for highly peaked distributions, and negative when the shape of the spectrum is flat relative to a Gaussian distribution (where kurtosis = 0). Kurtosis is often (but not necessarily) correlated with the spectral SD (Harrington, 2010). Here, kurtosis was calculated in *Praat* with the 'Get kurtosis' function (power = 2). The results are shown in Figure 4.10.

There was a significant main effect of facewear on kurtosis for /s/ [$F(7,42) = 3.66$, $p < .01$, $\eta_p^2 = .38$], /ʃ/ [$F(7,42) = 2.92$, $p < .05$, $\eta_p^2 = .33$], and /f/ [$F(7,42) = 6.11$, $p < .001$, $\eta_p^2 = .50$], but not for /θ/ ($p = .565$).

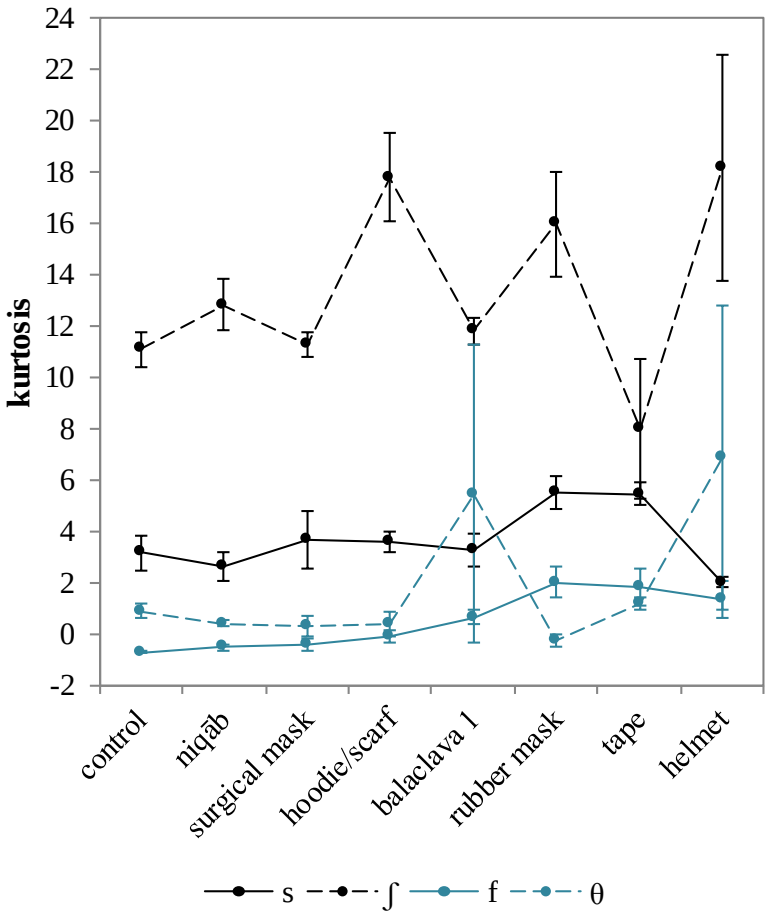

Figure 4.10. Kurtosis (dimensionless) for /s/, /ʃ/, /f/, and /θ/ produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. Note that the values on the y-axis start at –2 instead of zero. The error bars show the standard error of the mean.

The sibilant /ʃ/ obtained by far the highest kurtosis values (>10, except tape). This coincides with the low SD observed for /ʃ/. However, the results are again highly variable across facewear conditions. /ʃ/ obtained markedly higher values (compared to the baseline) when it was spoken through the scarf, rubber mask, and helmet. However, note the exceedingly large error bars in Figure 4.10 (especially for the helmet). Kurtosis for /s/ was much lower than for /ʃ/ (see low SD of /s/), but still positive (>2). The measures across facewear conditions were fairly consistent for /s/, with the exception of the rubber mask and tape, for which /s/ yielded noticeably (but not significantly) higher kurtosis values (i.e., more peaked distribution).

The non-sibilants had kurtosis values close to or below zero (with few exceptions). In view of their flat, diffuse spectra (with high SD), this was to be expected. In the

control condition, kurtosis was negative for /f/, and positive (but <2) for /θ/. Figure 4.10 illustrates that the patterns across facial disguise conditions were again quite consistent, except for the large increase in kurtosis for /θ/ produced in the balaclava and helmet conditions (note, however, the very large error bars). *Post-hoc* comparisons showed that kurtosis significantly differed from the baseline only when /θ/ was produced through the rubber mask ($p < .01$). For /f/, an increase in kurtosis was only evident in the rubber mask and tape conditions. Statistically, the values obtained for the balaclava and rubber mask significantly differed from the baseline ($ps < .01$).

4.2 Experiment 2: Voiceless plosives

Following the acoustic-phonetic analysis of voiceless fricatives, the second part of this chapter attends to intensity, spectral, and temporal measures of the voiceless plosives /p/, /t/, and /k/. The data were again extracted from the AVFC corpus (see Chapter 3). In the following sections, the most relevant acoustic characteristics of the plosives are detailed, and the methodology and results of the study are described.

4.2.1 Introduction

4.2.1.1 Aim and motivation

Following the acoustic-phonetic analysis of voiceless fricatives, the purpose of Experiment 2 is to examine acoustic properties of voiceless plosives, namely /p/, /t/, and /k/, and the extent to which these are modified when the plosives are produced while the talker's face is disguised by facewear. The two main motivations for analysing plosives are in accordance with the motives outlined for the fricative study.

Firstly, plosives are examined due to their relevance to forensic phonetic casework. The survey by Gold & French (2011) revealed that plosives, just like fricatives, are generally acknowledged by forensic speech scientists as an important speaker-discriminating parameter. English plosives obtained a mean Likert rating of 4.73 (*SD* = 1.49; see Gold & French, 2011: 301) when Gold & French's participants were asked to indicate how often they analyse plosives during casework (with '6' on a 6-point Likert scale denoting 'always'). Plosives were ranked second after fricatives, but their average Likert values closely matched those obtained for the fricatives.

Secondly, plosives exhibit an appreciable amount of distinctive acoustic energy in higher frequency bands of the acoustic spectrum. Llamas *et al.* (2008) noted that these acoustic characteristics make plosives just as susceptible to acoustic facewear effects as fricatives. Llamas *et al.* found that plosives (and fricatives) were the speech

sounds that were subject to the strongest acoustic filtering effects, especially when they were produced while the talker was wearing the balaclava and surgical mask. On this account, it seemed worthwhile to analyse plosives again on a larger scale.

The goal of this study is to examine whether and to what extent various forms of face coverings affect the acoustics of plosives. Once again, the focus is *not* on exploring or evaluating the acoustic properties which best discriminate between different places of articulation of plosives. Rather, this work has the aim of gaining insights into the effect of facewear on selected acoustic measures of each of /p/, /t/, and /k/. It is hypothesised that, due to their distinct acoustic structure (outlined below), the three plosives will be affected differently by facewear.[24]

The next section provides an overview of the most relevant articulatory and acoustic characteristics of /p/, /t/, and /k/. This is again followed by the description of the applied methodology, and the presentation of the statistical results.

4.2.1.2 /p t k/ revisited

Oral stop consonants can be described as a sequence of opening and closing events in the vocal tract, varying airflow patterns, and a chain of acoustic events ranging from absolute acoustic silence to high-energy explosions (Harrington, 2010). Owing to the aerodynamically and acoustically complex structure of plosives, this category of speech sounds is prone to a fairly large amount of variation within and between talkers (even of the same language/dialect; see e.g. Foulkes *et al.*, 2010).

From an articulatory point of view, oral stops are produced when a closure is formed somewhere along the vocal tract, which blocks the pulmonic egressive airstream from escaping the mouth (see segmentation boundary B1 in Figure 4.11). For

[24] The author acknowledges that differences between the terms ‘plosive’ and ‘stop (consonant)’ exist (Ladefoged & Maddieson, 1996). However, in the context of this thesis the terms are used interchangeably.

plosives produced at the alveolar place of articulation (like /t/), the occlusion is typically made with the tongue tip or blade pressing against the alveolar ridge. For velar stops (such as /k/) the tongue body pushes against the soft palate, and for bilabial plosives (like /p/) the upper and lower lip press against each other (Johnson, 2003; Roach, 2004). The constriction period ends when the closure of the articulators is released abruptly (see B2).

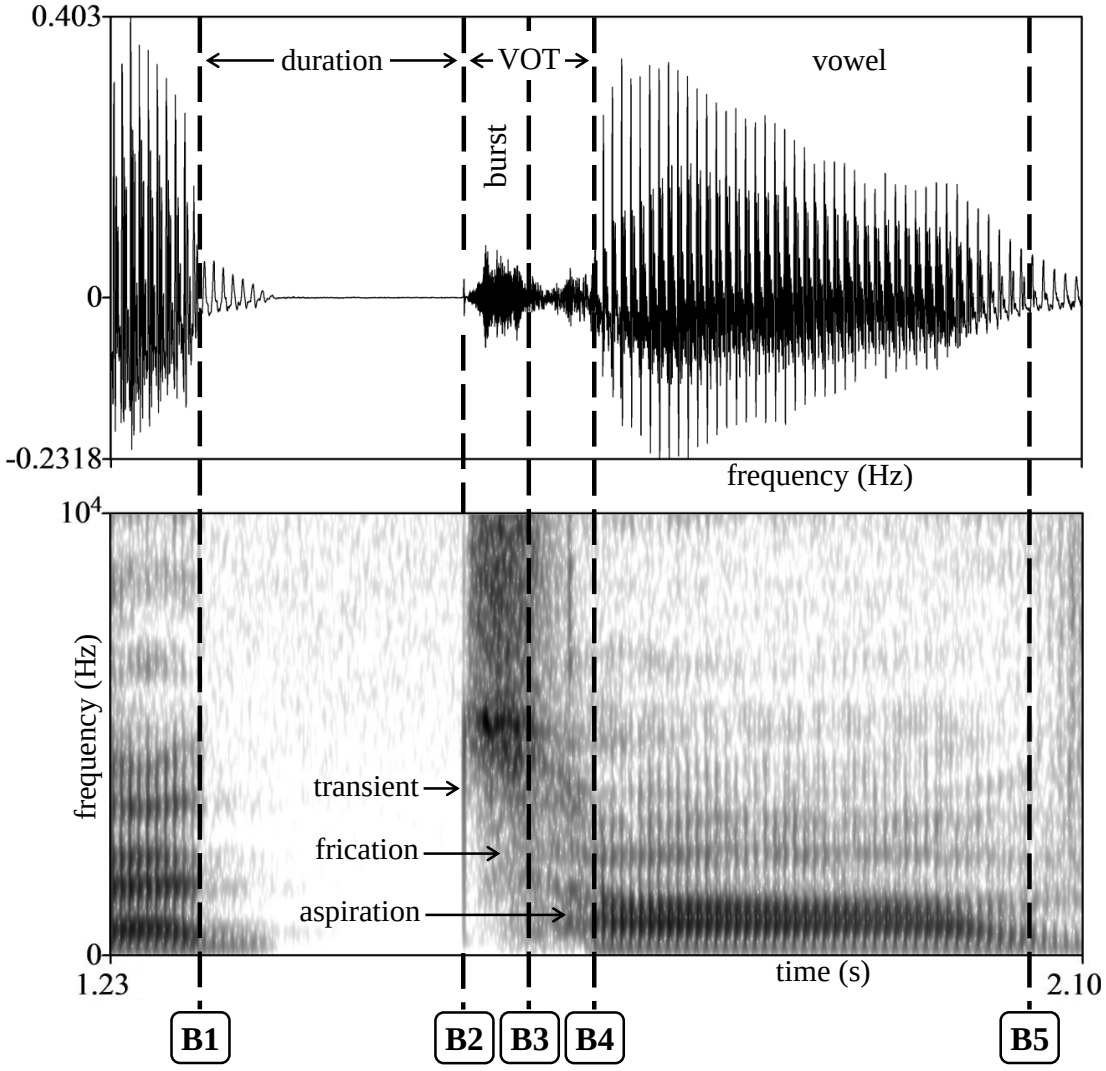

Figure 4.11. Pressure waveform (top) and wideband spectrogram (bottom) of [t^hɑ:] produced in syllable onset position by one of the male talkers recorded for the AVFC corpus (control condition). Boundary 1 ('B1') = beginning of plosive (onset of articulatory closure, acoustic near-silence); 'B2' = transient/beginning of frication (aperiodic energy created at closure release); 'B3' = beginning of aspiration (aperiodic energy created at glottis); 'B4' = beginning of voicing of adjacent voiced segment; 'B5' = end of voicing.

Acoustically, the constriction interval is defined as a period of silence. However, as Figure 4.11 shows, some acoustic energy (residual voicing) can still be present in this phase (Hayward, 2000; Machač & Skarnitzl, 2009; Harrington, 2010). This arises from the fact that the forming of the articulatory closure is not a sudden event, but occurs gradually as the articulators come together, and that the seal is not necessarily complete (Hayward, 2000). The acoustic (near-)silence typically ends with the release of air pressure at the 'burst'. The release phase is a sequence of 'transient', 'frication', and potentially 'aspiration' (Hayward, 2000; Harrington, 2010; Foulkes *et al.*, 2010). The transient corresponds to the moment of release of the articulators, and is visible in a pressure waveform as a vertical spike (see B2). Frication is the acoustic result of aperiodic energy created at closure release (see B2 to B3). Aspiration is aperiodic energy created at the glottis (see B3 to B4). Typically, the interval extending for approximately 20ms from the transient into the frication (and possibly aspiration) phase is referred to as the burst (Foulkes *et al.*, 2010; Harrington, 2010). In the current dataset (specified below), the average burst duration (measured by means of a Praat script in *Praat 5.3.24*) was 18ms for /p/ (*SD* = 14), 41ms for /t/ (*SD* = 17), and 20ms for /k/ (*SD* = 10). Please note that further information on the placement of segment boundaries will be given in §4.2.2.3.

When no aspiration follows the release of a voiceless plosive, the voiceless interval comes to an end at about the same time as the constriction interval (Hayward, 2000; Roach, 2004; Deterding & Nolan, 2007). If aspiration occurs – which is typically the case for spoken English – the voiceless interval extends beyond the constriction interval, and potentially overlaps with the formant transitions to the adjacent voiced segment (Hayward, 2000; Ladefoged & Disner, 2012). The formant transitions and differences in locus frequencies (i.e., the frequencies that a formant transition is heading towards; e.g. ~720Hz for labial, ~1.8kHz for alveolar, ~3kHz for velar combined with front vowels and below ~1kHz for velar combined with back vowels), are commonly consulted to distinguish the place of articulation of plosives (Delattre *et al.*, 1955; Hayward, 2000; Johnson, 2003; Harrington, 2010). In addition, there are place-dependent differences in the spectral shape of the release burst (Fant, 1960; Stevens, 1998; Harrington, 2010).

Figure 4.12 shows spectrograms of three isolated bursts produced by one of the male talkers represented in the current dataset (control condition). The leftmost spectrogram shows a typical /p/ burst. In general, (bi)labial plosives have rather faint (low-energy) bursts, which is predictable from the lack of a distinct front cavity during the production of (bi)labial sounds. The acoustic energy of a conventional /p/ burst is scattered over a wide frequency range (no distinct peaks), but is most commonly concentrated in lower frequencies, at around 0.5–1.5kHz. The latter gives rise to a falling spectral slope (Halle *et al.*, 1957; Johnson, 2003; Machač & Skarnitzl, 2009; Harrington, 2010; Ladefoged & Disner, 2012).

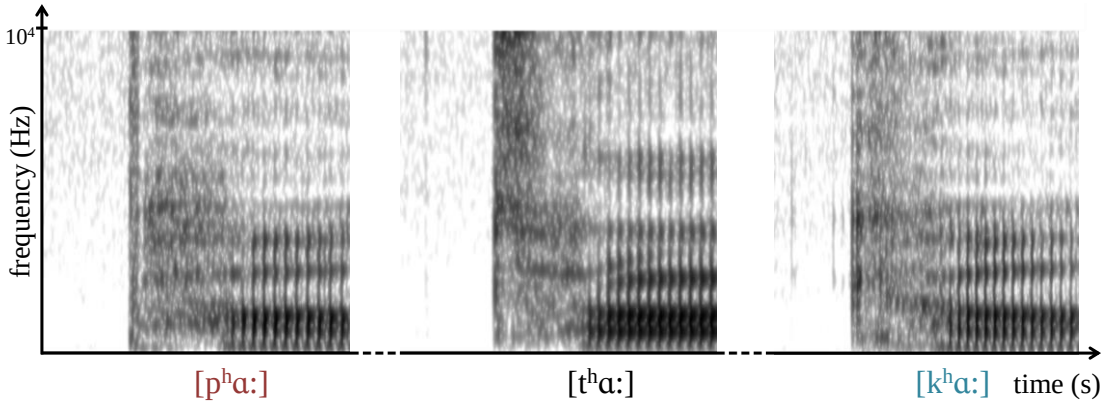

Figure 4.12. Wideband spectrogram of stop bursts of [p^h], [t^h], and [k^h] produced in syllable onset position (before /ɑ:/) by one of the male talkers recorded for the AVFC corpus (control condition). Note in particular the weak burst of /p/ (with energy concentrated in lower frequencies), the high-energy bursts of /t/ and /k/ (with high- and mid-frequency peaks), the non-continuous burst of /t/, and the multiple closure releases of /k/.

The middle spectrogram in Figure 4.12 illustrates the burst spectrum typically found for /t/. In comparison to (bi)labial stops, the spectra of alveolar plosives are generally flat. This means that the acoustic energy is fairly evenly distributed across the spectrum. However, alveolar stops sometimes exhibit energy maxima in low frequencies at ~500Hz, and high frequency bands above ~3–5kHz (formant-like structures, rising spectral slope). Moreover, alveolar plosives often feature non-continuous burst structures (Johnson, 2003; Machač & Skarnitzl, 2009; Harrington, 2010; Ladefoged & Disner, 2012).

Finally, the rightmost spectrogram in Figure 4.12 shows a typical /k/ burst. The burst spectra of velar stops are often characterised as ‘compact’, which indicates that the acoustic energy predominates in intermediate frequency bands. Velar plosives usually exhibit a mid-frequency peak in the 2–4kHz range (Harrington, 2010). Due to the long front cavity during the production of velar sounds, the intensity of stop bursts in velars is often higher than for speech sounds produced at more anterior places of articulation. Furthermore, velar plosives sometimes exhibit multiple bursts, which are usually thought to be caused by the Bernoulli Effect or ‘saliva sounds’ (Halle *et al.*, 1957; Johnson, 2003; Machač & Skarnitzl, 2009; Harrington, 2010; Ladefoged & Disner, 2012).

4.2.2 Method

4.2.2.1 Talkers and facewear

Following the methodology applied to the analysis of fricatives, six phonetically-trained, native British English speakers (three females, three males) were selected at random from the AVFC corpus. Their mean age was 25.6 years (*SD* = 6.2). Auditory evaluation of the speech material prior to the acoustic analysis showed that these six talkers did not pronounce the stop sounds in any unconventional way. The present experiment again included the control condition as a baseline, and the seven types of facewear that were tested in the fricative study (excluding the balaclava with the mouth hole, for the same reasons given in §4.1.2.1).

4.2.2.2 Speech material

The speech material for Experiment 2 consisted of three voiceless English stop consonants, namely /p/, /t/, and /k/. For each plosive, two tokens were again

randomly selected from the syllable onset and coda positions of each of the /C_1ɑ:C_2/ nonsense words recorded for the AVFC database. The selected material was again checked for reading and/or pronunciation errors, and samples were excluded and replaced by appropriate alternatives if any of these occurred. As per the fricative analysis, the audio recordings captured with the headband microphone were used.

In total, 576 plosive samples were selected and manually segmented: 6 talkers x 3 plosives x 2 syllable positions x 2 tokens x 1+7 facewear conditions (control + 7 types of facewear). In accordance with the fricative study (and the perception experiments reported in Chapter 5), only the results of the measurements taken from the syllable *onset* are reported.

4.2.2.3 Procedure

The plosives were hand-segmented in *Praat 5.3.24*. To ensure the consistent placement of boundaries, and to enhance segmentation accuracy, the segmentation points were carefully defined prior to the segmentation process. Foulkes *et al.* (2010) and Turk *et al.* (2006) point out that in spite of the fact that disputes exist, there is a general consensus across the phonetics community that both amplitude and spectrographic cues yield crucial segmental information. For this reason, both time-domain (waveform) and frequency-domain (spectrogram) representations of the speech signal should be used complementarily. The above authors recommend basing decisions on the waveform where possible, and consulting the spectrogram when in doubt.

Following these recommendations, the segment boundary positions were identified by visual inspection of the waveform for first-pass segmentation. Spectrograms were consulted for finer-grained decisions. All measurements, with the exception of the transient, were taken at zero crossings (see Foulkes *et al.*, 2010).

Segmentation boundary B1 (see Figure 4.11) denotes the beginning of the plosive. This point can be defined as the onset of the articulatory closure (visible as

diminution of periodic energy of the preceding voiced segment), or the moment of complete closure (acoustic silence or offset of voicing, which can persist for a short while after the onset of the oral constriction; see e.g. Turk *et al.*, 2006; Deterding & Nolan, 2007; Fuchs *et al.*, 2007; Foulkes *et al.*, 2010). In this study, the constriction onset criterion was applied. Accordingly, marker B1 was placed where periodicity in the spectrogram ended, and where a gradual attenuation of oscillations and a decrease in amplitude could be observed in the waveform.

Segmentation boundary B2 records the beginning of the burst. This is often (but not necessarily) visible as a sudden high-amplitude spike in the waveform (transient). In case of multiple release bursts, some researchers put the segment boundary on the final transient (e.g. Cho & Ladefoged, 1999), some on the initial (e.g. Lisker & Abramson, 1964; Turk *et al.*, 2006), and others on the most prominent one (e.g. Warner, 1996; Khattab & Al-Tamimi, 2008). Here, the visually most salient transient (i.e., the one with the highest amplitude) was chosen for analysis.

Segmentation boundary B3 marks the end of the burst, which corresponds to the end of frication and beginning of aspiration. This point can be difficult to determine, especially in noisy data. Typical cues in the waveform and spectrogram that support the adequate placement of this boundary are a sudden change in the spectrographic pattern, the onset of formant structures (corresponding to the formants of the adjacent voiced segment), or an abrupt drop in intensity, especially at lower frequencies (Foulkes *et al.*, 2010).

Segmentation boundary B4 denotes the beginning of voicing of the following vowel or voiced segment. The precise location of this point is disputed in the literature (see Foulkes *et al.*, 2010: 59, for an illustration). Lisker & Abramson (1964), for example, focus on the quasi-periodicity in the waveform (which reflects laryngeal vibration). Cho & Ladefoged (1999) look for the first complete vibration cycle of the vocal folds. Klatt (1975) refers to the onset of higher-energy striations in the second formant (F2) of the following voiced sound. In the present case, the F2 criterion by Klatt was applied (following e.g. Cho & McQueen, 2005; Deterding & Nolan, 2007; Fuchs *et al.*, 2007).

Lastly, segmentation boundary B5 indicates the end of voicing of the following voiced segment. Segmentation criteria were end of periodicity, F2 offset, and/or overall reduction in amplitude (Foulkes *et al.*, 2010).

The acoustic measurements were taken from wideband spectrograms (Gaussian; window length = 5ms) using a *Praat* script. It was again found after thorough consultation of the literature that the use of bandpass filters and pre-emphasis settings varies widely across studies. Some researchers apply high-pass filters with varying lower cut-off frequencies (e.g. 200Hz in Sundara, 2005, and Vicenik, 2010). This aims at removing the effects of pre-voicing or the air blast from the plosive release (Milenkovic, 1986). Following the fricative study, the analysed speech was unfiltered (no pre-emphasis filter) and sampled at 48kHz.

In sum, from each of /p/, /t/, and /k/, the following temporal, intensity, and spectral burst measures were taken:

A. Temporal measures
    - plosive closure duration (in milliseconds)
    - voice onset time (in milliseconds)
B. Intensity measure
    - relative burst intensity (in decibels)
C. Spectral measures
    - burst centre of gravity (in Hertz)
    - burst standard deviation (in Hertz)

4.2.3 Results

The statistical analysis of the data was performed by means of a series of two-way repeated-measures ANOVAs using *IBM SPSS Statistics V.19.0.0.1*. The dependent factors under consideration were ‘duration’, ‘voice onset time’ (VOT), ‘burst intensity’, and the first two statistical moments of the burst spectra, namely ‘centre of

gravity' (CG) and 'standard deviation' (SD). The independent within-subject factors were 'plosive' (/p/, /t/, /k/) and 'facewear' (control, balaclava without mouth hole, helmet, hoodie/scarf, *niqāb*, rubber mask, surgical mask, tape). There were again two between-subject factors, 'talker' and 'gender'. The results are reported in the form of averages across the speech elicited from all talkers. The effect of gender is dealt with in §4.2.3.1.

Effects are reported as significant when $p < .05$. Where Mauchly's test indicated that the assumption of sphericity had been violated, the degrees of freedom, *p*-values and effect sizes (η_p^2) were adjusted using the Greenhouse-Geisser correction (the correction factor ε is listed in the corresponding results table in such cases).

4.2.3.1 Overview

For a start, the data were statistically analysed after they had been averaged across plosive and facewear, respectively. This revealed that there was a significant main effect of plosive on all dependent factors, namely duration [$F(2,12) = 25.71$, $p < .001$, $\eta_p^2 = .81$], VOT [$F(2,12) = 23.00$, $p < .001$, $\eta_p^2 = .79$], and intensity [$F(2,12) = 104.23$, $p < .001$, $\eta_p^2 = .95$], as well as burst CG [$F(2,12) = 820.42$, $p < .000$, $\eta_p^2 = .99$] and SD [$F(2,12) = 90.85$, $p < .001$, $\eta_p^2 = .94$]. This finding implies that (averaged across control and facewear conditions) there were significant differences between the temporal, intensity, and spectral characteristics of /p/, /t/, and /k/.

The main effect of facewear was significant for duration [$F(7,42) = 7.38$, $p < .001$, $\eta_p^2 = .55$], VOT [$F(7,42) = 12.83$, $p < .001$, $\eta_p^2 = .68$], intensity [$F(7,42) = 6.46$, $p < .001$, $\eta_p^2 = .52$], CG [$F(7,42) = 8.31$, $p < .000$, $\eta_p^2 = .58$], and SD [$F(7,42) = 18.24$, $p < .001$, $\eta_p^2 = .75$]. This result indicates that (averaged across plosives) the various forms of facewear significantly altered the acoustic properties of the speech sounds.

Moreover, there was a significant interaction between plosive and facewear on intensity [$F(14,84) = 5.51$, $p < .001$, $\eta_p^2 = .48$], CG [$F(14,84) = 18.40$, $p < .001$, $\eta_p^2 = .75$], and SD [$F(14,84) = 7.84$, $p < .001$, $\eta_p^2 = .57$], but not on plosive duration ($p =$

.492) and VOT (p = .177). This means that the impact on intensity and spectral burst measures of each of the three plosives was dependent on the type of facial disguise condition the sounds had originally been produced in. That is, different plosives were differently affected by the facewear worn by the talker.

To explore the significant interactions in more depth, the data were subsequently analysed for each plosive individually. This ascertained the effect that each type of face mask had on the acoustic characteristics of each of /p/, /t/, and /k/.

As a final remark, a gender effect was found to act on the duration (p < .001), intensity, and CG (ps < .05) measures. Specifically, the effect was significant for duration in the tape (p < .01) and helmet (p < .05) conditions, for intensity in the control, rubber and surgical mask conditions (ps < .05), and for CG in the control, balaclava, helmet, hoodie/scarf, and surgical mask conditions (ps < .05). Despite these findings, and once more in acknowledgment of the fact that this is the less desirable approach, the decision was taken to average the results across female and male talkers (again, mainly for reasons of small sample sizes).

In the next sections, the results of Experiment 2 are presented for the temporal, intensity, and spectral moment measures separately. Again, the emphasis will be on the extent to which values obtained in the control condition differ from the corresponding values in each of the facewear conditions.

4.2.3.2 Plosive closure duration

The durations of the closure portions of the three plosives were measured from the closure (see B1 in Figure 4.11) to the release of the articulators (see B2), following e.g. Turk *et al.* (2006), Cho & McQueen (2005) and Stevens & Hajek (2004). The timestamps were retrieved with the 'Get starting point' function in *Praat*. The duration was then calculated by subtracting the timestamp of B1 from the timestamp of B2. The outcome of this procedure is shown in Figure 4.13.

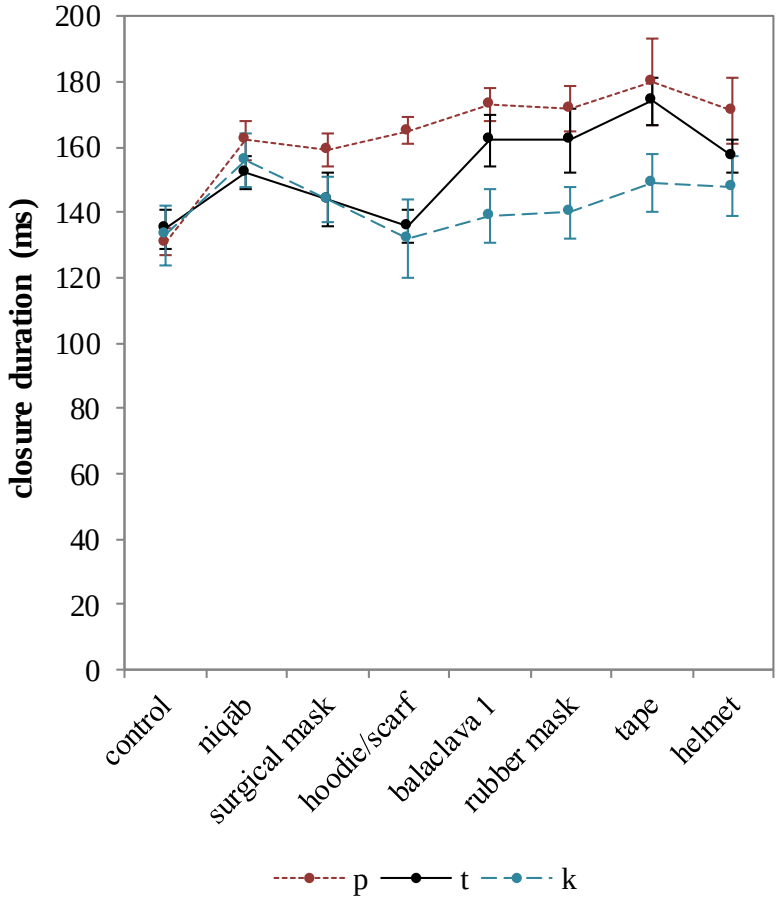

Figure 4.13. Mean plosive closure duration (in ms) of /p/, /t/, and /k/, produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. The error bars show the standard error of the mean.

The statistical analysis revealed a significant main effect of facewear on the stop closure durations of /p/ [$F(7,42) = 4.19$, $p < .01$, $\eta_p^2 = .41$] and /t/ [$F(7,42) = 3.75$, $p < .01$, $\eta_p^2 = .39$], but not of /k/ ($p = .582$). This implies that facewear on average significantly altered the duration of /p/ and /t/, but not the duration of /k/.

Figure 4.13 illustrates that the closure durations of all three plosives were very similar in the control condition. Differences between the three stops became apparent only in the facewear conditions. With very few exceptions, all duration values showed a trend to increase in facewear speech. The duration of /p/ and /t/ was in most facewear conditions up to 60ms longer than in the control condition. *Post-hoc* Bonferroni-adjusted pairwise comparisons revealed that compared to the baseline, the duration of /p/ was significantly longer when /p/ was produced in the surgical

mask ($p < .01$), balaclava, or hoodie/scarf ($ps < .05$) conditions. The duration of /t/ was significantly longer when /t/ was produced through the *niqāb* ($p < .05$).

4.2.3.3 Voice onset time

VOT specifies the timing relationship between the point of release of the stop closure (transient) and the phonation onset of the following vowel or voiced segment (Lisker & Abramson, 1964; Docherty, 1992; Cho & Ladefoged, 1999; Stevens & Hajek, 2004; Cho & McQueen, 2005). Here, VOT was computed in *Praat* by subtracting the timestamp retrieved for B2 from the timestamp for B4. The results can be seen in Figure 4.14.

The effect of facewear on VOT was significant for /p/ [$F(7,42) = 5.34$, $p < .001$, $\eta_p^2 = .47$] and /k/ [$F(7,42) = 10.12$, $p < .001$, $\eta_p^2 = .63$], but not for /t/ ($p = .114$). This indicates that facewear on average significantly changed VOT in /p/ and /k/, but not in /t/.

Figure 4.14 reveals that /t/ yields the highest VOT throughout (~120ms in control), while /p/ and /k/ each obtained lower VOT values across conditions (~90–100ms in control).[25] The patterns across the various facial disguise conditions are rather heterogeneous. However, as expected from the above statistical result, VOT of /t/ was relatively more stable across conditions (see the comparatively horizontal line for /t/ in Figure 4.14) than VOT of /p/ and /k/, respectively. Figure 4.14 illustrates that VOT of /p/ and /k/ increased in some conditions, especially the *niqāb*, hoodie/scarf, balaclava, and tape conditions. This contrasts with Llamas *et al.* (2008), who observed a slightly shorter VOT for /p/ when the stop was spoken through the balaclava (which possibly led to the observed misperception of /p/ as /b/).

[25] These VOT values are considerably higher than those expected for English. VOT of initial prevocalic stops in Southern British English provided by Docherty (1992: 116) are 42ms for /p/, 63ms for /t/, and 63ms for /k/. The high values in the present study may reflect the unnatural semantic environment and speaking style the plosives were elicited in (read nonsense syllables embedded in a controlled carrier phrase).

Statistically, only the VOT increase in /p/ caused by the tape yielded a significant result in *post-hoc* testing ($p < .05$).

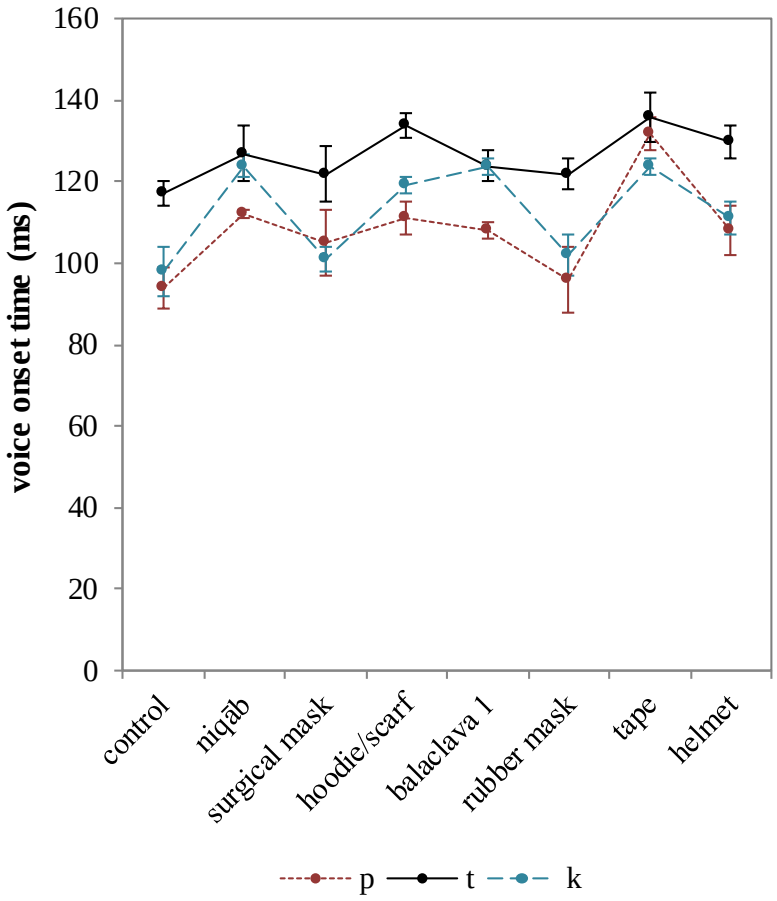

Figure 4.14. Mean voice onset time (in ms) of /p/, /t/, and /k/, produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. The error bars show the standard error of the mean.

4.2.3.4 Burst intensity

Following the presentation of the temporal measures, this section introduces the results of the burst intensity measure for each of /p/, /t/, and /k/. The burst intensity was calculated over the entire burst duration beginning at closure release (see B2) to aspiration onset (see B3). Different methods to compute the burst intensity have been proposed in the literature. For example, Fuchs *et al.* (2007) calculated the difference

between the intensity of the burst and the intensity of the following vowel midpoint. Colantoni & Marinescu (2008) subtracted the highest intensity value in the following vowel from the lowest intensity value in the plosive. Kirkham (2011) subtracted the peak intensity of the vowel from the intensity of the burst onset. In the present study, the 'relative intensity' of the three stops was obtained (see e.g. Stoel-Gammon *et al.*, 1994; Vicenik, 2010).

To do so, the maximum intensity of the burst and of the following vowel (in dB) was extracted using the 'To Intensity' (minimum pitch = 70Hz; time step = 0s; DC offset taken into account) and 'Get maximum' functions in *Praat*. Next, the burst intensity was calculated relative to the vowel intensity by subtracting the maximum intensity value of the burst from the maximum intensity value of the vowel. Consequently, the less prominent the burst, the larger would be the difference between the intensity of the burst and the vowel (and vice versa).

The results of this procedure are shown in Figure 4.15. Higher numerical values in the figure signify a weaker burst (higher relative intensity), and lower values denote a stronger burst (lower relative intensity). Note that the values on the y-axis were reversed in order to promote a more intuitive interpretation of the results. Data points towards the bottom of the graph now indicate a weaker burst than data points towards the top of the graph.

ANOVAs revealed a significant main effect of facewear on the relative burst intensity of /p/ [$F(7,42) = 5.05$, $p < .001$, $\eta_p^2 = .46$] and /t/ [$F(7,42) = 15.04$, $p < .001$, $\eta_p^2 = .72$]. The effect on the intensity of /k/ was not significant ($p = .241$). This suggests that facewear on average modified the intensity of the /p/ and /t/ bursts, but that the burst intensity of /k/ remained fairly stable across facewear conditions.

As expected, the bilabial plosive /p/ had the weakest burst (due to the largely missing front cavity during the production of /p/). This was also the case throughout the various facewear conditions, with the exception of the rubber mask condition. However, the relative burst intensity of /p/ decreased when /p/ was produced through facewear. This means that the burst became *more* intense relative to the vowel. This effect was most pronounced for the rubber mask. *Post-hoc* tests showed that the difference between the control and rubber mask samples was significant ($p < .01$).

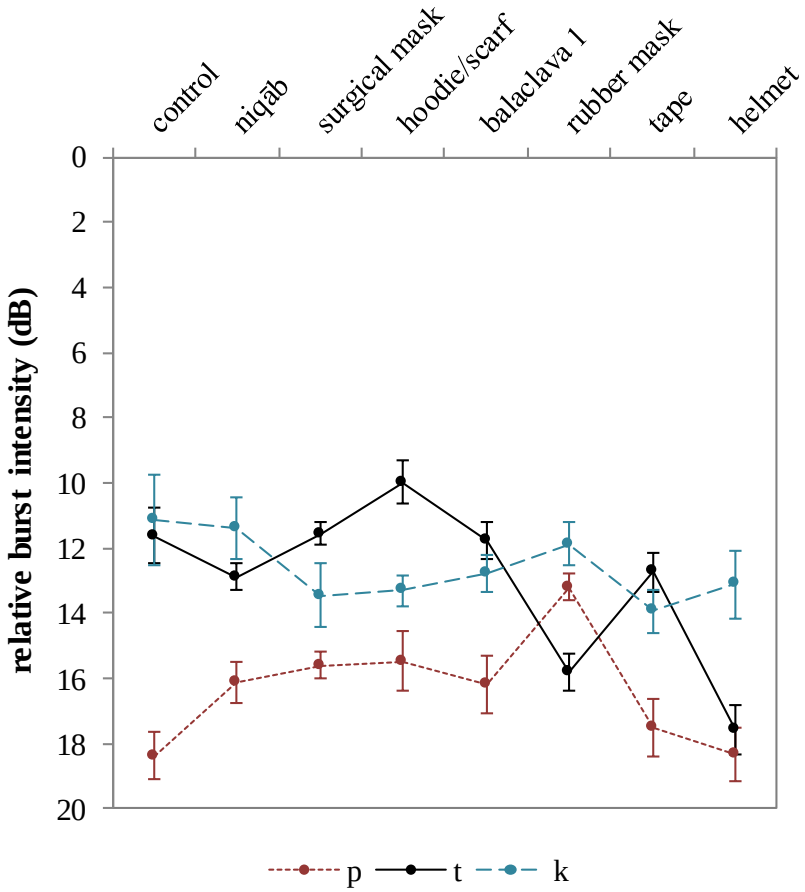

Figure 4.15. Mean relative burst intensity (in dB) of the bursts of /p/, /t/, and /k/, produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. Note that the values on the y-axis were reversed so as to facilitate a more intuitive interpretation of the data (data points towards the bottom of the graph denote a weak burst, while data points towards the top of the graph indicate a strong burst). The error bars show the standard error of the mean.

The alveolar stop /t/ had, again as anticipated, a lower relative intensity than /p/ in the control condition (i.e., a stronger burst). The results for /t/ across facewear conditions are fairly heterogeneous, with the relative burst intensity decreasing in some conditions (especially hoodie/scarf) and increasing in others. The latter was particularly noticeable in the rubber mask and helmet conditions, where the /t/ bursts were markedly less intense (relative to the vowel) than in the baseline. The intensity drop when /t/ was spoken through the *niqāb* accords with the results of Llamas *et al.* (2008), who allocated the misperception of /t/ as /p/ in the *niqāb* condition to changes to the burst intensity.

The relative burst intensity of /k/ equalled the relative burst intensity of /t/ in the control condition. However, the two stops were differently affected by the face masks. The statistical analysis suggested that facewear did not significantly alter the burst intensity of /k/. However, the patterns in Figure 4.15 give reason to believe that the /k/ burst was less intense (relative to the vowel) when it was produced while the talker's face was disguised.

4.2.3.5 Burst centre of gravity

In addition to the temporal and intensity measures, the first two spectral moments were computed from the power spectra derived from the burst noise. This was done in *Praat* following the same specifications given for the fricatives in §4.1.3.4 and §4.1.3.5. To recall, the centre of gravity (CG) is the mean frequency of the spectrum. The results of this analysis are plotted in Figure 4.16.

The statistical analysis of the data showed that the main effect of facewear on the CG was significant for all three stops, i.e., /p/ [$F(7,42) = 53.93$, $p < .001$, $\eta_p^2 = .90$], /t/ [$F(7,42) = 13.92$, $p < .001$, $\eta_p^2 = .70$], and /k/ [$F(7,42) = 2.25$, $p < .05$, $\eta_p^2 = .27$]). This implies that facewear significantly modified the centre frequency of each the /p/, /t/, and /k/ spectra.

In the control condition, the CG of the /p/ burst (at ~600Hz) was lower than the CG of the /k/ (at ~1.2kHz) and /t/ (at ~6kHz) bursts. This was again predictable from the short front cavity during the articulation of bilabial stops, and the compact spectrum where the energy is concentrated in low frequencies. With the exception of the balaclava and tape conditions, the CG of /p/ was very consistent in all facewear conditions (note the small errors bars in Figure 4.16). When /p/ was spoken through the balaclava, and in particular when it was produced with the talker's mouth taped shut, the CG increased, to around 1.5kHz (balaclava) and 3.5kHz (tape). This implies that more sound energy was now concentrated in higher frequency regions. *Post-hoc* tests showed that the effect of the tape was significant ($p < .01$).

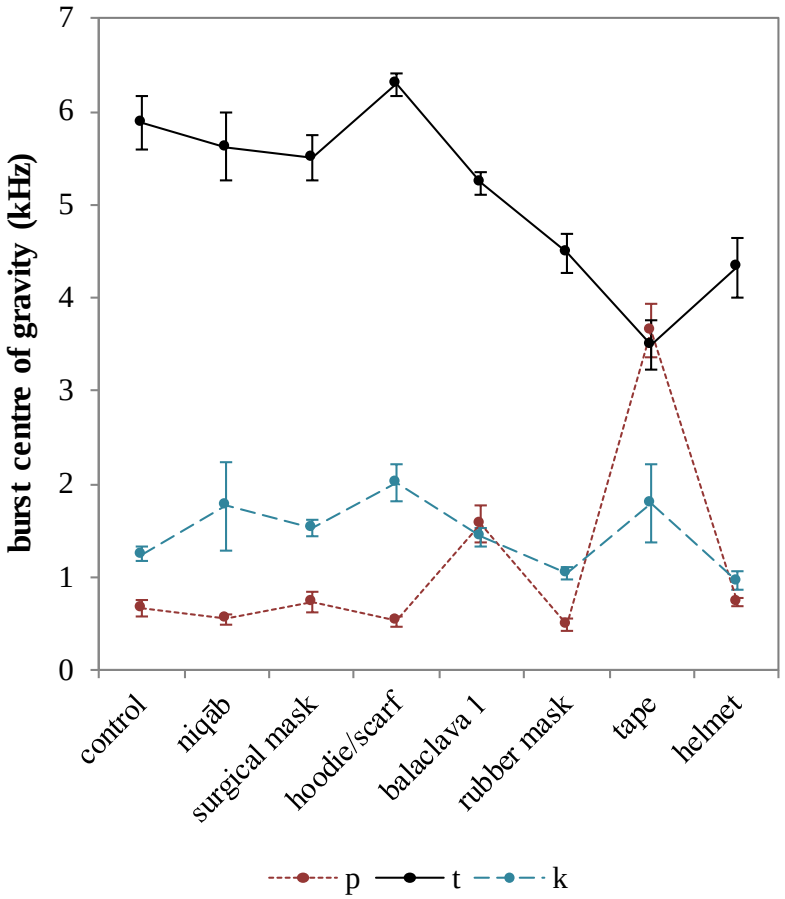

Figure 4.16. Mean centre of gravity (in kHz) of the burst of /p/, /t/, and /k/, produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. The error bars show the standard error of the mean.

Furthermore, /t/ revealed by far the highest burst CG of all plosives in the control samples (at ~6kHz). This again corroborates the literature. With one exception (hoodie/scarf), the centre frequency of the /t/ burst dropped in facewear speech. This effect was most prominent in the tape (CG reduced to ~3.5kHz), helmet (~4kHz), rubber mask (~4.5kHz), and balaclava (~5kHz) conditions. Statistically, the effect was significant in case of the tape ($p < .01$) and rubber mask ($p < .05$).

Lastly, Figure 4.16 demonstrates that the burst CG of /k/ in the baseline was rather low throughout the facewear conditions (~1–2kHz), but was higher than the CG of /p/. The CG measures across facewear conditions were relatively consistent, with facial masking causing a minor increase in the CG of /k/ in some conditions, and a slight decrease in others. It could be argued that the comparatively large error bars shown for some conditions in Figure 4.16 are to do with the production of multiple

release bursts by some talkers. These can give rise to multiple peaks in the spectrum, and might possibly shift the location of the centre frequency.

It was mentioned earlier (§4.2.2.3) that all experimental results are based on non-filtered speech. Some pilot experimentation using the 'Filter (pass Hann band)' and 'Filter (pre-emphasis)' functions in *Praat* demonstrated that different filter settings can induce a tremendous amount of variation in the spectral results. Specifically, a high-pass filter was applied to the present recordings, where the lower cut-off frequency was 200Hz (smoothing = 100Hz). This resulted in a rise of the burst CG of 19% (averaged across all plosives). When a pre-emphasis filter was applied (whereby spectral energy above 1kHz was enhanced by 6dB/octave), the CG increased by 165%. A combination of both filters led to an increase in CG of no less than 193% (for practical implications of these findings see §7.2).

4.2.3.6 Burst standard deviation

The standard deviation (SD) of the burst was computed for all three plosives following the specifications given in §4.1.3.5. As a reminder, the SD describes the dispersion of spectral energy around the centre frequency (CG). The outcome of the SD calculations is shown in Figure 4.17.

The main effect of facewear on the burst SD of /p/ [$F(7,42) = 31.09$, $p < .001$, $\eta_p^2 = .84$] and /k/ [$F(7,42) = 3.19$, $p < .01$, $\eta_p^2 = .35$]) was significant. The effect for /t/ was non-significant ($p = .940$), which suggests that the energy distribution around the CG of the /t/ burst did not significantly differ in facewear speech from that estimated from control speech.

In the control condition, the burst spectra of /p/ gave rise to the lowest SD of all three stops (~1–1.5kHz). When /p/ was spoken through facewear, the burst SD decreased (compared to the baseline) for some types of facial concealment (i.e., helmet and rubber mask, both SD < 1kHz), and increased for others. The SD increase was most

marked for the tape (SD > 3.5kHz) and balaclava (SD > 2kHz). The tape effect was significant ($p < .001$).

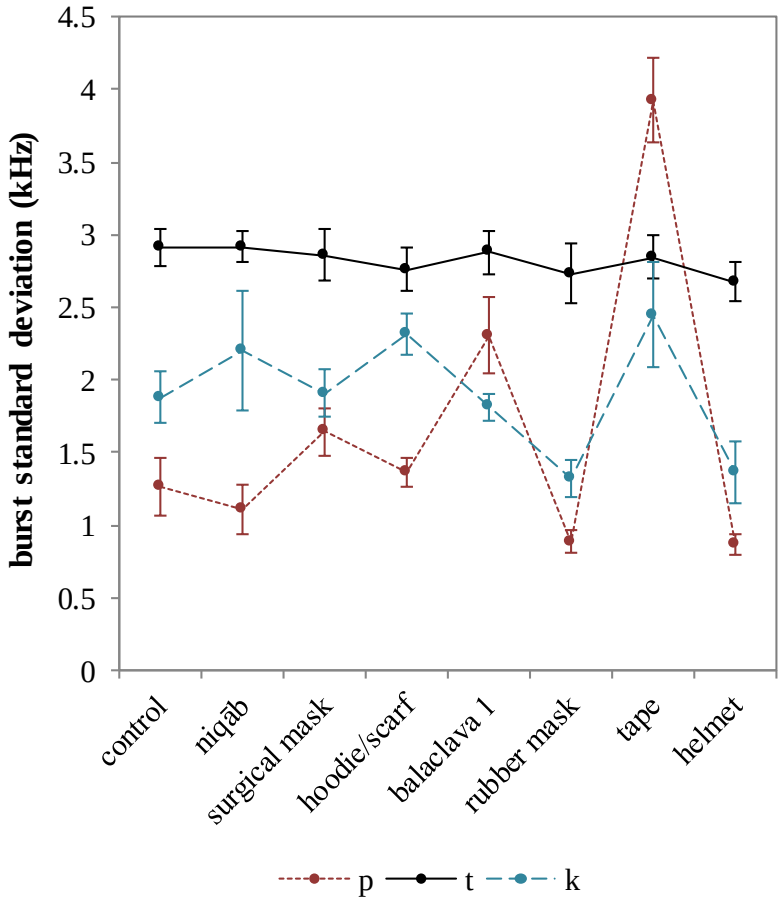

Figure 4.17. Mean standard deviation (in kHz) of the burst of /p/, /t/, and /k/, produced in syllable onset position, averaged across six talkers, for the control and each facewear condition separately. The error bars show the standard error of the mean.

The burst SD was highest for /t/, with values at around 3kHz in the control condition. This result was foreseeable from the diffuse spectra typically observed for /t/, where much of the acoustic energy is distributed across the entire range of the spectrum. The non-significant main effect of facewear reported above implies that the burst SD of /t/ was consistent across facewear conditions. The finding that facewear had no notable effect on the SD of /t/ is illustrated in Figure 4.17 by the relatively horizontal line.

As Figure 4.17 shows, the burst SD of /k/ (~2kHz) in the control condition lies between the SD of /p/ and /t/. The figure further illustrates that the SD results for /k/ are again quite variable (note the relatively large error bars in some instances). Most notably, the burst SD decreased when /k/ was spoken through the rubber mask and helmet, and increased in the hoodie/scarf and tape conditions.

Finally, experimentation with various filter and pre-emphasis settings (as per §4.2.3.5) suggested that the burst SD of the three plosives considerably varied when different filters or filter combinations were applied. For example, when the data (averaged across plosives) were high-pass filtered at 200Hz (smoothing = 100Hz), and/or a pre-emphasis filter (of 6dB/octave above 1kHz) was implemented, the burst SD increased by 10% (high-pass filtering), 119% (pre-emphasis), and 122% (both).

4.3 General discussion of Experiments 1 and 2

The third and last part of this chapter offers a general discussion of the results of Experiments 1 and 2. Both experiments focused on the acoustic-phonetic analysis of voiceless fricatives and plosives, which were produced while the talker's face/mouth was occluded by a balaclava (no mouth hole), hoodie/scarf combination, motorcycle helmet, *niqāb*, rubber mask, surgical mask, or a strip of tape; the balaclava with the mouth hole was excluded from the study. Fricatives and plosives were chosen for perceptual and acoustic reasons, and in consideration of their relevance as a consonantal feature commonly analysed by forensic speech scientists in casework. The high-quality audio recordings used for experimentation were obtained via the headband microphone placed approximately 2cm from the right-hand corner of the talker's mouth (see Chapter 3). Measurements were taken from non-filtered and non-pre-emphasised speech sampled at 48kHz. Spectral measures were taken from averaged, cepstrally-smoothed FFT-based power spectra computed from the medial phase of each fricative, and from the burst interval of each plosive.

In the first experiment, 768 fricatives were manually segmented following common phonetic conventions (6 talkers x 4 fricatives x 2 syllable positions x 2 tokens x 1+7 facewear conditions). The fricatives under investigation were two tokens each of the alveolar and post-alveolar sibilants (/s/ and /ʃ/, respectively), and the labiodental and dental non-sibilants (/f/ and /θ/, respectively), produced by six phonetically-trained native British English-speaking young adults. The acoustic measurements were based on the fricatives produced in syllable onset position (/C_1/ extracted from *He said /C_1a:C_2/.*). The following six measures, capturing the intensity and spectral properties of the frication noise, were taken into consideration: intensity (in dB), spectral peak (in Hz), centre of gravity (in Hz), standard deviation (in Hz), skewness (dimensionless), and kurtosis (dimensionless).

In the second experiment, 576 plosive samples were hand-segmented following carefully-defined segmentation guidelines (6 talkers x 3 plosives x 2 syllable positions x 2 tokens x 1+7 facewear conditions). The plosives of interest were two tokens each of the bilabial plosive /p/, the alveolar plosive /t/, and the velar plosive /k/, again produced by six phonetically-trained native British English speakers of the

same age group. In accordance with the fricative study, only the measurements taken from the plosives produced in syllable onset position were reported. The following temporal characteristics, as well as intensity and spectral properties of the burst, were analysed: plosive closure duration (in ms), voice onset time (in ms), relative burst intensity (in dB), burst centre of gravity (in Hz), and burst standard deviation (in Hz).

In the following sections, the main findings from both experiments are summarised by addressing the research questions raised at the beginning of this chapter. Moreover, the reader's attention is drawn to some of the observed qualitative acoustic changes to the speech sounds caused by facewear.

4.3.1 Acoustic facewear effects

The fundamental questions raised in this chapter were whether facewear changes the acoustic properties of voiceless fricatives and plosives, and specifically, whether temporal, intensity, and spectral properties of the sounds are modified to any extent when the segments are produced while the talker's face/mouth is disguised by a face covering. On the basis of the empirical work presented in this chapter, the short answer to this question is 'yes'.

For all four tested fricatives, a significant main effect of facewear on intensity and most spectral measures was observed. This demonstrates that the various types of face-concealing garments and headgear changed the acoustic structure of fricatives quite considerably. The main effect of facewear on the temporal, intensity, and spectral burst measures of plosives was statistically significant, which suggests that facewear on average significantly altered the acoustic properties of plosives.

More specifically, one goal of the study was to investigate whether the two classes of fricatives (sibilants and non-sibilants) were differently affected by facewear. It has been known for some time that the two classes differ in terms of their (articulatory and) acoustic characteristics, but that the acoustic properties *within* each class are fairly similar. In accordance with the literature, the sibilants and non-sibilants could

be distinguished (in the control condition) by their overall spectral shapes and energy distributions.[26] It follows that facewear affected the two classes to different degrees and in different manners.

Firstly, the spectral peaks and centres of gravity of the sibilants (especially /ʃ/) were overall quite consistent across facewear conditions, while the peaks and centres of the non-sibilants were considerably modified by facewear (however, there was also a high amount of variation across samples, especially in the peak data). The spectral centres of gravity of the non-sibilants tended to decrease in facewear speech.

Secondly, except for an increase in some conditions (especially tape, helmet, and rubber mask), the standard deviations of both the sibilants and non-sibilants were only marginally altered by facewear. However, there was more variation in the non-sibilant data altogether (higher standard errors).

Thirdly, when facewear was involved, the spectral distribution of the sibilants was relatively stable, with the exception of marked positive skewing (towards lower frequencies) especially for /s/ produced through the rubber mask, balaclava, and tape, and negative skewing in the hoodie/scarf and rubber mask conditions for /ʃ/. The skewness values for the non-sibilants were generally subject to more variation, and increased with facewear.

Fourthly, the spectral distributions tended to become more peaked (higher kurtosis) for /ʃ/ (especially in the hoodie/scarf, rubber mask, and helmet conditions), but were reasonably consistent for /s/ in facewear speech (except rubber mask and tape). Kurtosis was relatively consistent for the non-sibilants in some of the conditions, but

[26] The sibilants were characterised by greater intensity than the non-sibilants (/s/ = ~65dB, /ʃ/ = ~63dB), and much of their spectral energy was accumulated in distinct spectral peaks (/s/ = ~7–9kHz, /ʃ/ = ~3kHz) and centres of gravity (/s/ = ~8kHz, /ʃ/ = ~4kHz) with a low standard deviation (/s/ = ~2.2kHz, /ʃ/ = ~1.8kHz). Regarding their spectral shapes, /s/ revealed peaked spectra (low but positive kurtosis; gradual rise to the peak) and skewing towards higher frequencies (positive but low skewness), while /ʃ/ showed highly peaked spectra (high kurtosis; steep slope to the peak) and skewing towards lower frequencies (high skewness). The non-sibilants were specified by lower intensity (/f/ = ~50dB, /θ/ = ~55dB), and diffuse, flat spectra (low/negative skewness and kurtosis) with highly variable peaks, no major resonances or regions of prominence, and a large standard deviation (~4–5.5kHz; but energy by trend concentrated in higher frequencies, at ~6–9kHz for /f/, and ~7–10kHz for /θ/).

showed a tendency to increase in the facewear conditions (especially for the balaclava, helmet, rubber mask, and tape).

Considering next the acoustics of plosives, it was found that /p/, /t/, and /k/ could be distinguished from one another (in the control condition) based on place-dependent differences in the intensity and spectral shape of the release burst, and to some extent from the temporal patterns (VOT, but not closure duration).[27] Owing to these distinctive acoustic structures, it is again unsurprising that the acoustic properties of /p/, /t/, and /k/ were modified differently by facewear.

Firstly, the intensity of the weak /p/ burst tended to increase (especially for the rubber mask), the intensity of the strong /t/ burst decreased in some conditions (especially rubber mask and helmet) and increased in others (especially hoodie/scarf), and the strong /k/ burst tended to weaken in facewear speech.

Secondly, the low centre of gravity of the /p/ burst was stable across the various facial disguise conditions (except balaclava and tape), the high centre of gravity of the /t/ burst tended to drop (especially for the tape, helmet, rubber mask, and balaclava), and the mid-frequency centre of gravity of /k/ was relatively consistent.

Thirdly, the low burst standard deviation of /p/ showed highly variable patterns (highest for the tape and balaclava, lowest for the rubber mask and helmet), the high burst standard deviation of /t/ was consistent, and the intermediate burst standard deviation of /k/ was again prone to a lot of variation across facewear conditions (lowest for rubber mask and helmet).

Fourthly, there was more often an increase in plosive closure duration, and in part of VOT, than there was a reduction. Although absent in the baseline, major differences in duration became apparent only in the facewear conditions. The durations of /p/

[27] In accordance with the literature, /p/ was characterised by low-energy bursts (~18dB relative/~55dB absolute intensity) with an energy concentration in low frequency bands (~600Hz), and a low standard deviation (1–1.5kHz). The bursts of both /t/ and /k/ exhibited high amounts of acoustic energy (~11dB relative/62dB absolute intensity). However, energy in the /t/ burst dominated in high frequencies, at ~6kHz (but SD = ~3kHz), while the energy of the /k/ burst was accumulated in intermediate bands, at ~1.2kHz (SD = ~2kHz). The closure durations of all three stops were similar (~130ms). VOT of /t/ (~120–130ms) was longer than VOT of /p/ and of /k/ (both ~90–100ms).

and /t/ were up to 60ms longer when the plosives were spoken through facewear (especially the tape, helmet, balaclava, rubber mask, and surgical mask conditions). The long VOT of /t/ in the baseline was relatively more stable than the VOT values of either /p/ or /k/, which exhibited a larger amount of variation across conditions than /t/, and by trend increased in facewear speech (especially in the tape, *niqāb*, hoodie/scarf, and balaclava conditions).

4.3.2 Acoustic absorption and speech compensation

Based on the findings from this study, which type of face covering can be put forward as having the most detrimental effect on the acoustics of fricatives and plosives? The answer to this question is not as straightforward. As pointed out repeatedly, the extent to which facewear modified the acoustic characteristics of fricatives and plosives was largely dependent on the specific type of sound tested (see the significant facewear x fricative/plosive type interactions). Put another way, different face masks appear to alter the acoustic-phonetic properties of each individual fricative and plosive differently.

Furthermore, there was a fairly large amount of variation across and within facewear conditions (see the large error bars), and the facewear-induced acoustic changes were not always statistically significant (see the non-significant *post-hoc* tests). The latter may in part be the consequence of the comparatively small sample sizes. Nevertheless, the present data offer a good estimate of the kinds of acoustic modifications that one should expect when working with facewear speech.

On the whole, the smallest acoustic effect on the plosive burst and fricative spectra was observed when the speech was produced through the *niqāb* and surgical mask. Compared to the baseline (no facewear) condition, the intensity of the frication noise was barely affected when the fricatives were spoken through these two types of facewear. This was predictable from the relatively thin and lightweight textiles of both coverings, which were not expected to absorb sound energy to a great extent

(for a list of facewear materials see §3.1.2). Some minor changes to the spectral peak and moments were observed, but these were only prominent, if at all, for the non-sibilants (which are prone to a lot of variation in any case, especially /θ/).

The findings for the *niqāb* and surgical mask disguise were generally confirmed for the plosives. The spectral burst properties were little affected when /p/, /t/, and /k/ were spoken through either of the two face coverings. However, marginal changes to the intensity of the burst noise were observed (increase for /p/, reduction for /t/ and /k/). Also, the temporal measures were modified to some degree, in that closure duration and VOT tended to increase.

As for the hoodie/scarf and balaclava (no mouth hole) guises, the results were overall more heterogeneous. Most strikingly, when the fricatives were produced through the scarf, the intensity of the frication noise increased (compared to the baseline). This was only a minor effect. However, it seems a counterintuitive one at first, considering that the scarf material was thicker and heavier than the materials of the *niqāb* and surgical mask, for which virtually no increase in intensity was noted. It can be speculated that the talkers may have actively compensated for wearing the facewear by speaking more loudly. This strategy of increasing the level of vocal effort may have counterbalanced the perceptual effects of sound energy absorption caused by the mask materials. In other words, the consequences of raising the loudness of the voice may have 'outweighed' some of the acoustic filtering effects of the mask material. This (deliberate or automatic) articulatory compensation behaviour may also explain the increase in burst intensity when the plosives were spoken through the scarf or balaclava, and specifically, the intensification of the (weak) /p/ and (strong) /t/ burst.

The above observations are consistent with the literature presented in §2.3, and in particular with the results of the transmission loss experiment conducted by Llamas *et al.* (2008). Llamas and colleagues found that the surgical mask (thin layers of pleated paper) inhibited sound transmission to a greater extent than the ostensibly more sound-absorbing fabrics of the balaclava (knitted acrylic) and the two tested scarves (knitted wool/acrylic mix, knitted polyester). Hence, it can be concluded that heavier, thicker, or more densely-woven fabrics and materials do not necessarily

change the acoustics of facewear speech to a greater extent than thinner, lighter, or more porous ones. Rather, the acoustic facewear effects on the speech signal appear to be the consequence of a combined effect of acoustic transmission loss and of active modifications to the talker's speaking behaviour.

The largest impact on the acoustic structure of fricatives and plosives occurred for the helmet, tape, and rubber mask disguises. As expected, the effect of facewear on the intensity of the frication noise (especially of non-sibilants) was most noticeable for the helmet, tape, and rubber mask. Altogether, the changes in intensity triggered by the three masks were less prominent in the plosive data, but they were still noticeable to some degree. In reference to the case made above, the talkers may still have compensated for having their face/mouth covered by speaking more loudly. However, the mask materials (especially the solid, highly sound-absorbing shell of the helmet) presumably filtered out and attenuated acoustic energy much more heavily than was the case in the other facewear conditions.

In terms of the spectral properties of fricatives (especially non-sibilants), the centre of gravity significantly decreased (by ~1–2kHz) when the speech was produced through the helmet, tape, and rubber mask (i.e., there was a more positively-skewed spectral distribution). The centre of gravity of the /t/ burst decreased in these conditions (and also for the balaclava), but increased (along with the standard deviation) for /p/ in the balaclava and tape conditions.

Lastly, the helmet, tape and rubber mask gave rise to significant temporal modifications to the plosives. There is a relatively high level of variation in the data, and the results are only subtle (in the range of 30–40ms). Nonetheless, the changes to the temporal composition of the speech may be interpreted as a more prolonged pronunciation on the part of the talkers, and hence as another indication that the talkers adjusted some of their articulatory habits.

4.3.3 Sound energy migration

The aim of this section is to point the reader towards some of the qualitative acoustic characteristics of facewear speech that were noted by the author while working with the data. Some of the most common observations are illustrated in Figure 4.18. The figure shows spectrograms (left) and spectra (right) for /s/ produced (before /ɑː/) in the control and helmet conditions by three male talkers extracted from the pool of talkers in the fricative study (labelled talkers 'A', 'B' and 'C'). For illustrative purposes, each /s/ production (and a portion of the adjacent /ɑː/) was isolated from the /C_1ɑːC_2/ syllable that it was originally produced in. After that, spectrograms and spectra were computed in *Praat*. Within each spectrogram shown in Figure 4.18, the left-hand side shows /s/ spoken in the control condition, and the right-hand side shows /s/ produced through the motorcycle helmet (chosen for its marked acoustic effects). Within the spectral displays, the black line represents the control condition, and the red line denotes the helmet condition.

First of all, the speech samples extracted for all three talkers demonstrate that facewear reduces the intensity of the frication noise. This effect becomes evident in the lighter shading (less blackening) of the 'helmet speech' in the spectrograms, and in the consistently lower red line (at least above a certain threshold) in the corresponding spectra.

Next, facewear sometimes brought about the shaping or intensification of formant-like patterns in the signal. These are exemplified in Figure 4.18 for /s/ produced through the helmet by talkers B and C. For talker C, the frequencies particularly around 1.5kHz were amplified (marked by the arrows). For talker B, the formants (again indicated by the arrows) appear to be the result of attenuation of acoustic energy surrounding the bandwidth of the formant(s) (especially between 5.5kHz and 7.5kHz), rather than enhancement of certain frequencies in the spectrum. Either way, such formant-like structures will give rise to additional peaks in the spectrum. This in turn may significantly alter the location of the centre of gravity and other spectral measures.

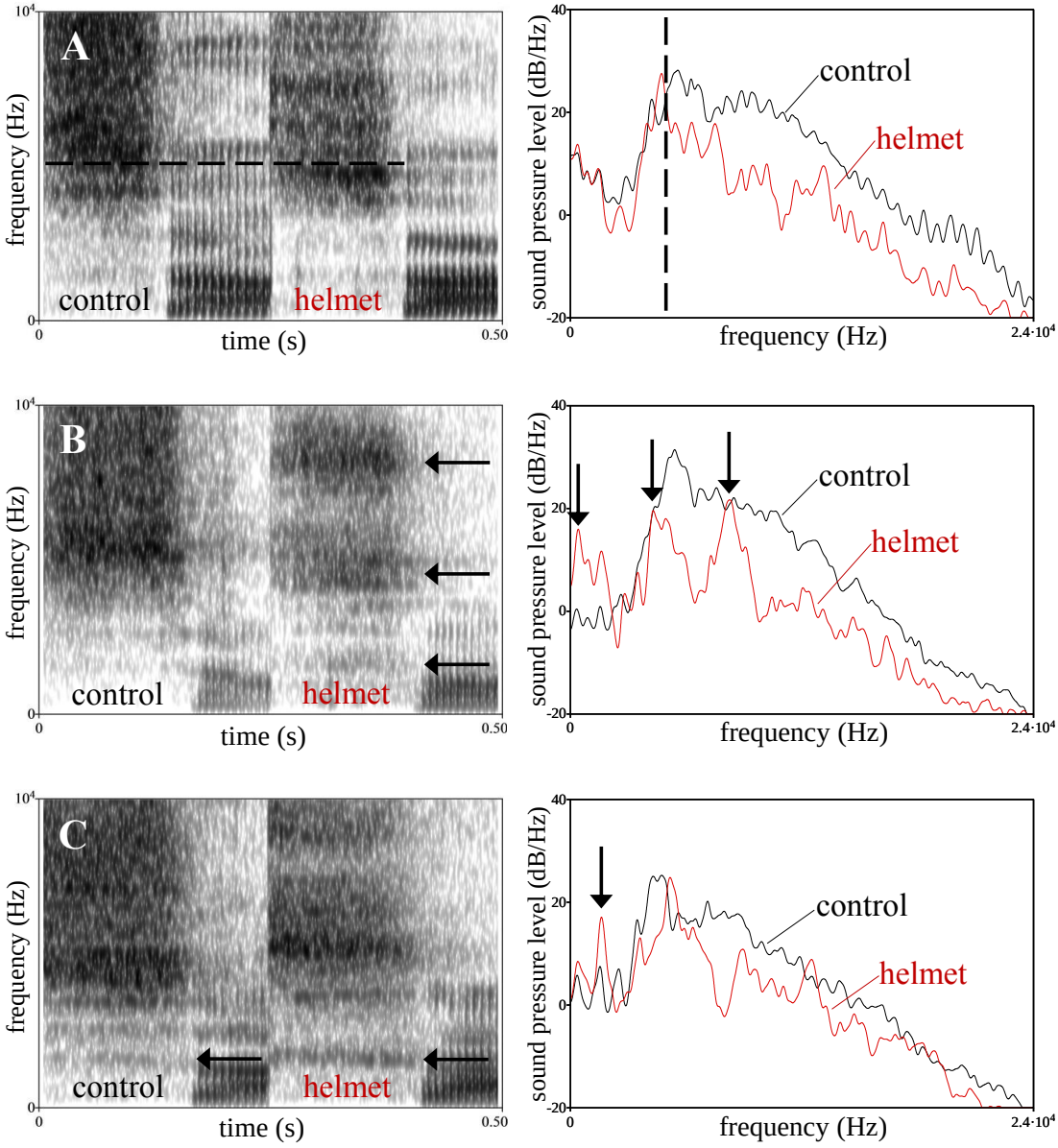

Figure 4.18. Modifications to the intensity and spectral shapes of frication noise as an artefact of facewear speech. The wideband spectrograms (left, top to bottom) and cepstrally-smoothed power spectra (right, top to bottom) each show /s/ spoken in syllable onset position (before /ɑː/) by three of the male talkers recorded for the AVFC corpus (labelled 'A', 'B', and 'C'). The figure illustrates some of the facewear-induced qualitative changes to speech that were typically observed in the present data (here, using the example of speech produced through the helmet).

Furthermore, and most importantly, the current data affirm the assumption brought forward by Llamas *et al.* (2008) that acoustic energy, especially in higher frequency bands, will be attenuated or filtered out when the speech is produced while the talker's face is concealed by a mask. In Figure 4.18, this effect is best illustrated by

talker A. As can be seen in the spectrogram, the (approximate) threshold above which acoustic energy is filtered out of the signal lies at 5–6kHz (denoted by the horizontal black, dashed line). In the spectral display, the black (control) and red (helmet) lines overlap (or approximate each other) up to this point, and then diverge.

These observations suggest that acoustic energy in facewear speech is damped above the approximate threshold of 5–6kHz, and that energy below the threshold is less (or not at all) attenuated by comparison. In other words, less sound energy is concentrated in higher frequency bands in facewear speech than in control speech, and relatively more energy in lower bands. Concerning the spectral properties of facewear speech, this means that as a result of these acoustic filtering effects, the centre frequencies will be 'artificially' lowered, and the spectral distribution will be positively skewed. Following the terminology used by Stanton *et al.* (1988), this relative increase of acoustic energy in lower frequency ranges at the expense of energy in higher bands can be described as 'energy migration' (see also §2.3.3).

On a conceptual level, the artificial shift of centre frequencies caused by acoustic filtering resembles the well-established 'telephone effect' in forensic speech science. The telephone effect refers to the fact that certain frequencies of the acoustic speech signal will be attenuated or filtered out when the speech is transmitted via a telephone channel. Research has shown that this will ultimately distort certain acoustic-phonetic measurements, especially of the first formant of high vowels. The effect was first described by Künzel (2001) for landline telephony, and later analysed in mobile phone transmission, e.g. by Byrne & Foulkes (2004), and VoIP (voice communication over Internet Protocol), e.g. by Fecher (2008).

Künzel (2001: 82f.) specifically suggests that the centre frequency of the first formant of high vowels is artificially shifted upwards, firstly because the formant bandwidth will be reduced from the bottom (lower cut-off frequency for landline in Germany, where the study took place, was 400–500Hz at the time). Secondly, the relative weight of the higher harmonics of the formant will be increased if they fall within the slope of the transmission channel. Arguably, a conceptually similar process occurs for facewear speech. If we consider facewear to act as a low-pass filter, which suppresses acoustic energy in particular above 5–6kHz, we are in a

better position to explain some of the modifications to the speech spectra (such as the typical downward shift of the centre of gravity).

The cut-off frequencies and passband slopes of the 'facewear filters' are, of course, not nearly as clear-cut as they are for landline and mobile telephony, where the filters and speech codecs implemented are highly standardised (Fecher, 2008). Further research – e.g. in the style of Llamas *et al.*'s 2008 transmission loss experiment – will be necessary to establish the transmission slopes for different fabrics and materials, and to understand the acoustic facewear effects in their entirety. But even at this early stage, the findings from the present study clearly suggest that (forensic) speech analysts give consideration to the above observations when working with speech recordings where it is known or suspected that the speech was produced while the talker's face was concealed by a face covering of some sort.

4.3.4 Summary

In conclusion, the study presented in this chapter provides experimental data showing that facewear has the potential to considerably modify certain intensity, temporal, and spectral characteristics of voiceless fricatives and plosives. The main findings can be summed up as follows:

- the non-sibilant fricatives /f/ and /θ/ are acoustically more affected by facewear, exhibit more variation (in means) across facewear conditions, and are subject to more variability across samples, than are the sibilants /s/ and /ʃ/ (see Table 4.1)
- facewear affects the voiceless plosives /p/, /t/, and /k/ differently (see Table 4.2)
- facewear effects on the acoustic speech signal are the consequence of a combined effect of a) acoustic transmission loss caused by the mask material, and b) active changes to a talker's articulatory behaviour
- transmission loss and energy migration
    - acoustic absorption particularly in higher frequency bands (above 5–6kHz)

    - centre of gravity shifts in frication noise and burst spectra (1–2kHz lower)

- deliberate/automatic articulatory compensations for speech perturbations
    - raised vocal effort: increase of intensity (loudness) of frication/burst noise
    - more prolonged/exaggerated pronunciation: e.g. up to 60ms longer plosive closure durations, and by trend longer voice onset times

- facewear-induced modifications to the acoustic-phonetic properties of voiceless fricatives and plosives vary greatly with facewear type
    - most noticeable acoustic effects in the motorcycle helmet, tape, and rubber mask conditions (thick, heavy, sound-absorbing materials)
    - minor effects for the *niqāb* and surgical mask (thin, lightweight textiles)
    - however, heavier, thicker, or more densely-woven fabrics and materials do not necessarily change the acoustics of facewear speech to a greater extent than thinner, lighter, or more porous ones (see e.g. intensity increase despite transmission loss in the hoodie/scarf and balaclava conditions)

voiceless fricatives (sibilants /s ʃ/, non-sibilants /f θ/)		
centre of gravity	/s ʃ/	consistent across facewear conditions (except minor rise for /ʃ/, tape)
	/f θ/	mostly lower (up to ~2kHz; esp. balaclava, rubber mask, tape, helmet)
standard deviation	/s ʃ/	consistent (higher for /ʃ/, tape; and for /s/, tape, helmet)
	/f θ/	quite consistent, but more variation across samples
skewness	/s ʃ/	variable (lower: for /s/, esp. rubber mask, balaclava, tape; and for /ʃ/, tape; higher: for /ʃ/, esp. hoodie/scarf, rubber mask)
	/f θ/	variable, but by trend higher (esp. balaclava, tape, helmet, rubber mask)
kurtosis	/s ʃ/	quite consistent for /s/ (except higher for rubber mask, tape); variable for /ʃ/ (higher: esp. hoodie/scarf, rubber mask, helmet; lower: tape)
	/f θ/	quite consistent, but by trend higher (for /ʃ/, esp. balaclava, helmet; for /s/, esp. rubber mask, tape); high variation across samples overall
intensity	/s ʃ/	variable (minor effect: *niqāb* , surgical mask; slightly higher: hoodie/scarf, balaclava; significantly lower: rubber mask, helmet, tape)
	/f θ/	as per sibilants /s ʃ/

Table 4.1. Summary of the main findings from the spectral peak, centre of gravity, standard deviation, skewness, kurtosis, and intensity measurements of the four voiceless fricatives /s/, /ʃ/, /f/, and /θ/ (Experiment 1).

voiceless plosives /p t k/		
burst centre of gravity	/p/	consistent except for marked rise for balaclava and tape
	/t/	mostly lower (esp. tape, helmet, rubber mask, balaclava)
	/k/	quite consistent
burst standard deviation	/p/	variable (lower: esp. rubber mask, helmet; higher: esp. tape, balaclava)
	/t/	quite consistent
	/k/	variable (lower: esp. rubber mask, helmet; higher: esp. tape)
plosive closure duration	/p/	up to 60ms longer across facewear conditions
	/t/	up to 60ms longer across facewear conditions
	/k/	quite consistent, but by trend longer (esp. *niqāb*)
voice onset time	/p/	variable, but by trend longer (esp. tape)
	/t/	quite consistent, but by trend longer (esp. hoodie/scarf, tape)
	/k/	variable, but by trend longer (esp. *niqāb* , hoodie/scarf, balaclava, tape)
burst intensity	/p/	mostly higher (esp. rubber mask)
	/t/	variable (lower: esp. rubber mask, helmet; higher: esp. hoodie/scarf)
	/k/	by trend lower

Table 4.2. Summary of the main findings from the burst centre of gravity, burst standard deviation, plosive closure duration, voice onset time, and burst intensity measurements of the three voiceless plosives /p/, /t/, and /k/ (Experiment 2).

Following from the acoustic-phonetic study of selected speech sounds in the present chapter, the next chapter shifts the focus from the acoustic characteristics of consonants to the perceptual properties of consonants spoken through facewear. Chapter 5 presents two speech perception experiments, both of which examine the ability of phonetically-untrained listeners to auditorily and auditory-visually identify consonants which were produced while the talker's face was disguised by facewear.

5 Auditory-visual perception of facewear speech

5.1 Introduction

The current chapter presents two speech perception experiments which deal with the auditory-visual perception of speech produced through facewear, and more explicitly, with the identification of consonants embedded in CVC syllables. Altogether, the experiments address the following questions:

- Does facewear change the perceptual properties of spoken English consonants? Specifically, is the identification of consonants hindered when the consonants have been produced while the talker's face is disguised by facewear?
- Do lay listeners more accurately identify the consonants when they can watch the talker's articulating face and hear the talker's voice (compared to only hear the talker's voice)?
- Assuming that auditory-visual facewear effects emerge, can listeners extract visual speech information from the talker's face even when the face is partly or fully concealed by a face covering?

Before moving on to describing the methodology and results of the study, the reader is familiarised with the research area of auditory-visual speech processing. The focus will be on previous studies which have attempted to identify the facial regions that are most informative to the observer during auditory-only, auditory-visual, and visual-only speech processing.[28] It is demonstrated that facewear research greatly contributes to this line of research (as previously argued in Fecher & Watt, 2013).[29]

[28] On a terminological note, despite the fact that conceptual differences may exist to some, the terms 'recognition' and 'identification' are used interchangeably in this context.

[29] Some of the results of this study were presented in 2012 at the *British Association of Academic Phoneticians (BAAP) Colloquium*, the *21st Annual Conference of the International Association of Forensic Phonetics and Acoustics (IAFPA)*, the *32nd Australasian Experimental Psychology Conference*, and *'The Social Side of Speech' conference* (MARCS Institute, Sydney), and in 2013 at the *12th International Conference on Auditory-Visual Speech Processing (AVSP)* and the *Postgraduate and Academic Researchers in Linguistics at York (PARLAY)* conference.

5.1.1 Auditory-visual (AV) speech processing

5.1.1.1 Multimodality of speech processing

Humans perceive their surrounding environment in a multimodal way. The capacity of the brain to integrate input from different modalities has been acknowledged as an important aspect of the human perceptual system. Speech has indeed been described as the prototypical case of multimodal perception, which is apprehended by visual (speechreading), auditory (hearing), and even haptic (touch) means (Bernstein *et al.*, 2000; Massaro, 2001; Grant, 2003; Swerts & Krahmer, 2008; Gick & Derrick, 2009; Ito *et al.*, 2009). In fact, auditory-visual (AV) speech has even been termed the *primary* mode of speech perception (Rosenblum, 2005).

Since the early work by Cotton (1935), Sumby & Pollack (1954), Fisher (1968), Greenberg & Bode (1968), and others, it has been extensively demonstrated that speech intelligibility is better maintained when both auditory and facial cues generated during speech production are available to the perceiver. The linguistic information derived from the acoustic signal and the visible speech gestures from the talker's articulating face have been shown to combine into a coherent percept, which may be more richly specified than that obtained from either of the unimodal sources alone. The widely-studied ventriloquist and McGurk effects have often been cited as evidence of the automaticity of multimodal integration. In these studies, mismatched (incongruent) auditory and visual speech stimuli are presented synchronously. The emerging 'fusion illusions' (Vatikiotis-Bateson *et al.*, 1998: 937) demonstrate in a most striking way the extent to which visual information from the face can influence auditory speech perception (see also McGurk & MacDonald, 1976; Massaro, 1987; Benoît *et al.*, 1996; Massaro, 1998; Thomas & Jordan, 2002; Burnham & Dodd, 2004; Tiippana *et al.*, 2004; Brungart & Simpson, 2005; Rosenblum, 2005; Hazan *et al.*, 2006; Rosenblum *et al.*, 2007; Hazan & Li, 2008; Kroos & Dreves, 2008; Sekiyama & Burnham, 2008; Chen & Hazan, 2009; Davis & Kim, 2009; Fitzpatrick & Kim, 2010; Hazan *et al.*, 2010).

The complementary nature of speech perception has often been investigated by testing how speech cues missing in one channel can be recovered from the other

channel, in cases where either the auditory or the visual information from the talker's face is disrupted or lost from the signal (Grant & Seitz, 2000; Grant, 2003).

Firstly, studies have shown that listeners rely more heavily (or exclusively) upon visual information when *acoustic* speech cues are absent or distorted by additive noise (Sumby & Pollack, 1954; Summerfield, 1987; Marassa & Lansing, 1995; Rosenblum & Saldaña, 1996; Preminger *et al.*, 1998; Grant & Seitz, 2000; Thomas & Jordan, 2002; Grant, 2003; Munhall *et al.*, 2004; Schwartz *et al.*, 2004; Thomas & Jordan, 2004; Rosenblum, 2005; Tuomainen *et al.*, 2005; Davis & Kim, 2006; Hazan *et al.*, 2006; Lidestam & Beskow, 2006; Hazan *et al.*, 2008; Swerts & Krahmer, 2008; Kim *et al.*, 2009; Hazan *et al.*, 2010; Stephens & Holt, 2010; Jordan & Thomas, 2011). In noisy or reverberant environments, important acoustic attributes of the signal, which are relevant for the identification of phonetic units, can be weak or distorted. This can cause considerable ambiguity in the auditory channel, and thus impair speech perception. Visual speech information can help to restore the missing auditory speech cues (even when a portion of the auditory signal is absent).[30]

Secondly, the interaction of auditory and visual cues during AV speech processing has been examined when the *image* accompanying the auditory stimulus is partially or wholly obscured (Greenberg & Bode, 1968; Marassa & Lansing, 1995; Rosenblum & Saldaña, 1996; Preminger *et al.*, 1998; Thomas & Jordan, 2004; Davis & Kim, 2006; Swerts & Krahmer, 2008; Kim *et al.*, 2009; Jordan & Thomas, 2011). Among other things, studies revealed that the cognitive processes responsible for the perception of facial movement during AV speech perception are notably resistant to loss of coarse (configural) information. This loss can arise, for example, from the reduced physical size of the talking face caused by increased distance between observer and image (Erber, 1974; Jordan & Sergeant, 2000), from facial inversion (Rosenblum *et al.*, 2000), changes to the horizontal viewing angle (Jordan &

[30] In the AV phonemic restoration experiment by Shahin & Miller (2009), participants listened to tri-syllabic words while a portion of each word was artificially replaced by white noise. They then judged whether the utterances sounded continuous or interrupted. Phonemic restoration occurred even when the noise durations were quite long. Fagel (2005) found that participants in his study perceived audible speech when lip movements were presented along with acoustic noise, despite the complete absence of an auditory speech signal.

Thomas, 2001) and facial orientation (Jordan & Bevan, 1997), or the removal of colour from the facial image (Jordan *et al.*, 2000; Thomas & Jordan, 2004).

Furthermore, researchers have shown that AV speech perception is fairly resilient to the loss of fine facial detail arising from the reduction of the image quality on the featural level. For example, previous research demonstrated that the visual contribution to speech perception does not require images with a high spatial resolution (high clarity). Before the era of digital video processing, blurred images for experimentation were created by placing transparent screens or other objects between the speaker and listener. In one of the earliest studies, Stone (1957) investigated how the degree of facial exposure, facial expression and lip mobility affects speechreading performance by placing plastic screens with different-sized openings in front of the talker during filming. Greenberg & Bode (1968) studied consonant recognition for full-face compared to lips-only exposure by means of an opaque mask that was positioned over a television monitor to obscure all of the talker's face but the lips, mandible and larynx. Berger *et al.* (1971) obtained two facial exposure conditions by using translucent fiber-filled theatrical face masks which exposed different parts of the face of the talker, who was positioned behind a glass window. Erber (1979) placed rough-surfaced plexiglass between lipreaders and the talker, and increased the distance between the two so as to gradually increase the amount of blurring. Nowadays, researchers such as Munhall *et al.* (2004), Thomas & Jordan (2004, 2002), or Jordan & Sergeant (2000), are using ever more sophisticated video capture and post-processing techniques (e.g. digital band-pass filters) to produce the desired effects (see e.g. Figure 5.1).

Figure 5.1. Facial images showing the talker in the study by Munhall *et al.* (2004) under various viewing conditions. In all band-pass filtered conditions (except the rightmost) an improvement of speech intelligibility in noise (keyword recognition) was found, but no filtered version reached the accuracy level of the unfiltered video (leftmost). Reproduced with adaptation from Munhall *et al.* (2004: 577).

5.1.1.2 In search of visual speech cues

The experimental techniques, the tested linguistic material, and the region(s) of interest in the talker's face vary widely across previous research on auditory-visual speech processing and speechreading (visual-only speech perception). However, one common goal of the studies has been to identify the facial areas which are most informative to the observer. Suprasegmental (prosodic) information in an utterance was found to be recovered largely on the basis of movements in the upper part of the face (eyes, eyebrows) and head motion (Summerfield, 1987; Lansing & McConkie, 1994; Munhall *et al.*, 2004; Davis & Kim, 2006; Swerts & Krahmer, 2006, 2008; Cvejic *et al.*, 2010, 2011). Segmental information (that concerning consonants and vowels) was shown to be mainly encoded in the lower part of the face. It is the latter that will be of further interest in the present context.

Linguistically-relevant visual events that encode segmental information are primarily located in the mouth region. This is of course plausible when recalling the principal role of the lips during speech production. Many of the early studies suggest that oral movement alone provides all of the segmental speech cues available in a fully-visible talking face. Summerfield (1979), for example, presented facial displays in which the talker's lips were coated with fluorescent make-up so that only the lips could be seen. He found that these lips-only displays produced a significant increase in sentence comprehension in noise compared to auditory-only presentation of the stimuli. IJsseldijk (1992) reports that word, phrase, and sentence identification only marginally improves when the AV stimuli (i.e., simultaneous audio + video) consisted of full-face as opposed to mouth-only displays. Using more refined video processing techniques, Thomas & Jordan (2004) systematically varied the amount of dynamic and static facial information visible to the observer by digitally modifying narrowly-defined areas of the talkers' face. The mouth region was thereby defined as an area within 2mm of the border of the lips (see Figure 5.2). Contrary to earlier results, they found that observers were still able to extract useful information for AV speech identification from the outer mouth region even when the mouth itself was static in or absent from the image.

Figure 5.2. Static examples of the video displays used by Thomas & Jordan (2004). The mouth and eyes+nose were either absent (4), present (2 + 3), or both (1), or the 'facial frame' and the eye+nose were absent (8), present (7 + 6), or both (5). Visual and auditory-visual speech recognition increased even when the displays only showed the talker's extraoral movements. Reproduced with adaptation from Thomas & Jordan (2004: 879).

There exists by now a wealth of evidence which suggests that facial cues other than those provided by the lips are important during AV speech processing, for example, visual information conveyed from inside the mouth. The results from studies are mixed, but generally point towards an involvement of the tongue and teeth (e.g. Badin *et al.*, 2010). Erber (1974), for instance, reports that visual word recognition improves as the illumination of the posterior surface of the tongue is intensified. The 'point-light' study by Rosenblum *et al.* (1996) revealed that markers positioned on the tongue and teeth enhanced speech recognition.[31] By contrast, Preminger *et al.* (1998) selectively masked certain facial features (e.g. the tongue or lips) by selecting the corresponding pixels and setting them all to the same grey level, which effectively eliminated the selected features from view. They found that the visibility of the tongue and teeth were only of limited importance during speechreading.

It has additionally been found that the mandible (lower jaw) is an important component of visual speech perception. The movement of the jaw is closely coordinated with the movement of other articulators (e.g. the tongue). It therefore mirrors the vocal tract changes that lead to consonant- and vowel-specific constrictions along the vocal tract, and hence supports visual speech intelligibility (Marassa & Lansing, 1995; Yehia *et al.*, 1998; Vatikiotis-Bateson & Ostry, 1999; Thomas & Jordan, 2004). Marassa & Lansing (1995) experimentally limited facial

[31] Point-light studies test observers' speech perception performance when the observers are presented only with the kinematic information from reflective markers that are strategically placed on the talker's otherwise darkened face.

movement outside the lip and mandible region. They found no significant differences in speechreading performance between the condition where the whole face moved, and where only the lips and mandible moved. By contrast, Rosenblum *et al.* (1996) showed that speech recognition did not further improve when 'point-light' markers were added to the jaw (or chin, forehead, and nose, for that matter).

Finally, speech production involves the finely coordinated motion of oral and extraoral facial muscles (Lesner, 1988). The muscle contractions necessary to control articulatory movement are in some sense 'imprinted' on the facial surface, including the chin and cheeks. Greenberg & Bode (1968) tested consonant recognition under the condition that participants were exposed to the full face, or to the talker's lips, mandible, and larynx only. They observed an advantage of seeing the entire face over seeing the mouth and neck region alone. Scheinberg (1980) suggests that observers use the rapid cheek movements (inflating of the cheeks) as a perceptual cue to discriminate between consonants that look similar in the mouth (visemes). This observation was affirmed by Preminger *et al.* (1998), who showed that movement of the oral articulators is highly correlated with the rapid movement of the extraoral areas. They proposed the existence of four locations of major 'jitter': the chin, the cheeks at the sides of the mouth, puffing of the face and cheeks near the upper lip, and the sides of the nose (Preminger *et al.*, 1998: 570). According to the authors, facial speech cues located at the chin and cheeks are sufficient for identifying a range of visemes. For example, the cheeks appear to be useful for identifying the plosive /p/ and the affricate /tʃ/, while cheek puffing and chin wrinkling (caused by muscle contractions used to raise the lower lip) seem to be helpful for recognising the fricative /f/. These findings accord with those of Lidestam & Beskow (2006), who found that consonant, word and sentence identification was more accurate when observers were presented with an image of a human talker as opposed to an animated talking head. The authors argue that subtle phonemic features (like cheek inflation during the pronunciation of a bilabial stop but not a homorganic nasal) are available from the image of the human talker but not from the avatar.

5.1.1.3 Towards more natural facial occlusion

Over the years, several different techniques have been applied to determine the facial speech cues that are most informative for the observer. These include eye-tracking (Lansing & McConkie, 1994; Munhall & Vatikiotis-Bateson, 1998; Vatikiotis-Bateson *et al.*, 1998; Paré *et al.*, 2003) and motion capture using 'point-light' displays (Rosenblum *et al.*, 1996; Rosenblum & Saldaña, 1996; Jordan *et al.*, 2000). Another common method is selective visual masking by means of the 'window technique' (Marassa & Lansing, 1995; Preminger *et al.*, 1998; Thomas & Jordan, 2004; Davis & Kim, 2006; Swerts & Krahmer, 2008; Jordan & Thomas, 2011). Here, different orofacial areas are systematically eliminated from view, and the effect on speech recognition is tested. Some displays from previous work are shown in Figures 5.3 to 5.5.

Figure 5.3. Still images of the video stimuli used in research on selective visual masking during speechreading by Preminger *et al.* (1998). The talker is shown under five different masking conditions, namely no masking, tongue+teeth, mouth, mouth+above, and mouth+below masking (left to right). Reproduced with adaptation from Preminger *et al.* (1998: 566 and 571).

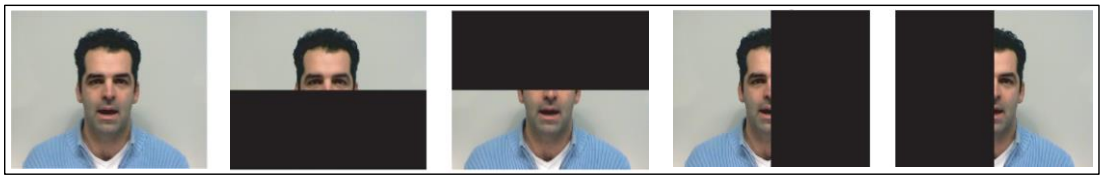

Figure 5.4. Static images representing different versions of the video stimuli applied in the study on perceptual processing of facial markers of prominence by Swerts & Krahmer (2008). The horizontal and vertical bars superimposed with the images blacken out the upper, lower, left or right side of the talker's face. Reproduced with adaptation from Swerts & Krahmer (2008: 229).

Figure 5.5. Facial displays used in the experiments on the effects of facial occlusion on visual and auditory-visual speech perception by Jordan & Thomas (2011). Various parts of the talker's face were occluded by vertical, horizontal, or diagonal black polygons added to the images in post-production. Reproduced with adaptation from Jordan & Thomas (2011: 2276).

The window technique allows the investigation of orofacial structures independently of one another, and also the evaluation of the relative prominence of one region in a talker's face over another. Despite these benefits, the method has recently faced criticism. Jordan & Thomas (2011: 2271) contend that selective masking of a talker's face could unintentionally induce 'an abnormal focus of visual and attentional resources that may exaggerate the feature's influence on visual speech perception and distort an understanding of the influence of other areas of the face'. They further remark that from a cognitive point of view, the salience of the unoccluded area (i.e., the part of the face which remains visible) could unintentionally be heightened because observers know that the display shows only that area throughout the experimental trial. Alternatively, observers may be encouraged to focus their gaze or attention on the occluded area (the distracting object) rather than the visible facial regions (the same argument is brought forward by Marassa & Lansing, 1995). Another point of criticism raised by Thomas & Jordan (2004) concerns the role of holistic facial information during AV speech processing, which could be underestimated in this case. The holistic perception of a face (i.e., the perception of the face as a complete entity rather than a set of individual facial features) has been shown to play a role not only during face recognition (see e.g. Frowd *et al.*, 2012), but also during AV speech recognition (see e.g. the facial inversion studies by Jordan & Bevan, 1997, or Rosenblum *et al.*, 2000).

Building on these arguments, Jordan & Thomas (2011: 2271) encourage researchers to make use of more *realistic* occlusions when setting out to explore the extent to which observers tolerate loss of perceptual information that is brought about by facial occlusion.

> [T]he tolerance of visual speech perception towards loss of facial information is far from understood. In particular, while previous studies have focused on maintaining information from an individual facial feature (e.g., lips or mouth), a natural system of visual and audiovisual speech perception is likely to develop to cope with everyday occlusions that do not obscure all of a face except for the precise parameters of a particular feature.

Jordan & Thomas (2011: 2271) further argue that the stimuli used in relevant studies should occlude the talker's face in a more *natural* way, in order to reflect the fact that

> faces in everyday environments are naturally obscured simply and extensively in various uncontrolled ways, by intervening objects, other people, shadows, the talker's own hand or hair, and so on.

Undoubtedly, one such category of realistic facial occlusions that meets the criteria of realism and naturalness is the set of various types of face-concealing garments and headgear which are the object of investigation in this thesis.[32]

5.1.2 Aim of the study

In the majority of previous studies on the effects of facial occlusion on auditory-visual speech processing, the talker's facial appearance was experimentally modified

[32] The importance of using natural facial images for testing is also stressed by Heath & Moore (2011). They found that a face disguised with a balaclava had caused the face overshadowing effect, but that a blank disk, which was covering the talker's face in a less natural way, had not provoked the effect (see also §7.3). Furthermore, the only published work (that the author is aware of) which has previously looked at other forms of 'natural impoverishment' of the visual speech signal, and how they affect speech processing, is that by Fuchs *et al.* (2010) and Kitano *et al.* (1985). These researchers have tested the effects of facial hair (moustaches, beards) on AV speech intelligibility in noise (Fuchs *et al.*, 2010) and on speechreading performance by hearing-impaired participants (Kitano *et al.*, 1985). Facial hair can cover articulatorily-important parts of the face (lips, teeth, larynx), for which reason an influence on visual speech processing seems plausible. Whereas Kitano *et al.* did not find any significant effects, Fuchs and colleagues observed a trend towards reduced speech intelligibility in the moustache condition.

in video post-production. For example, selected parts of the talker's face were blackened out (window technique), or the pixels corresponding to the movement of major articulators (e.g. lips, tongue) and pre-defined orofacial regions (e.g. chin, cheeks) were adjusted in a way that the movement of certain facial areas were eliminated from the experimental display.

The present study makes use of the footage recorded for the AVFC corpus (see Chapter 3) to test the effects of facial concealment on AV speech perception. Hence, the approach differs from the procedures in preceding studies in two main ways. Firstly, the facial displays are not 'artificially' modified by post-processing the image in any way, but the talker's face is at the time of recording 'naturally' disguised by a range of face coverings which (more or less) commonly occur in everyday spoken communication situations. Secondly, the test material is arguably more natural and realistic from an acoustic/auditory point of view. This arises from the fact that the study takes the modifications to the acoustic speech signal caused by facewear into account. These may result from the acoustic absorption on the part of the mask material, and/or the adaptations to the talker's speech productions (see §2.1.2.2). The study thus tests the *combined* perceptual effect of the facewear-induced changes to speech production and acoustics, and of the impoverished facial image.

The goal of the study principally follows that of Llamas *et al.* (2008), which is to ascertain whether speech intelligibility is adversely affected when the talker's face is disguised by a face covering. However, consonant identification is examined on a much larger scale (more talkers, more face coverings, etc.), and the methodology and data analysis employed is extended and refined a great deal.[33]

[33] The study addresses some of the weak points in Llamas *et al.* (2008), which are to do with the methodology and lack of a statistical analysis. In Llamas *et al.* (2008) only two talkers were recorded (thus giving no account of intra-talker variation), and the participants' native language and gender were not sufficiently balanced. Moreover, no distinction was made between the identification of consonants presented in onset or coda position of the tested CVC words, listeners could predict the identity of a word based on the words presented in preceding trials, and responses were elicited with handwritten response sheets (the latter may have drawn attention away from the computer screen used for stimuli prompting).

5.2 Experiment 3: Auditory-visual consonant identification in quiet listening conditions

This part of Chapter 5 discusses the methodology and results of Experiment 3, which tests the ability of phonetically-untrained listeners to identify syllable-onset consonants. These were produced under different facewear conditions, and are presented in auditory-only (AO) or auditory-visual (AV) formats. This aims at investigating the impact of various forms of facial occlusion on AO and AV consonant perception under (otherwise) optimal listening and viewing conditions. Experiment 3 establishes a baseline which facilitates comparison with the results from a subsequent speech-in-noise experiment (Experiment 4, presented in §5.3).

5.2.1 Method

5.2.1.1 Participants

Forty-four native English-speaking students (26 females, 18 males) were recruited at the University of York, United Kingdom. Their mean age was 19.5 years (*SD* = 1.5). None of them reported a history of hearing impairment, and all had normal or corrected-to-normal vision. No participant reported previous experience of wearing any type of facewear, or interacting with people who do so, on a regular basis. All volunteers participated in the experiment in return for a small remuneration.

5.2.1.2 Speech material

The speech material was extracted from the AVFC corpus presented in Chapter 3. Of the three simultaneous continuous audio recordings made during each recording session, this experiment used those captured with the DPA 4066 Omnidirectional

Headband Microphone placed at approximately 2cm from the right-hand corner of each talker's mouth (48kHz, 768kbit/s, 16-bit signed integer PCM encoding). The selected audio was not normalised for amplitude in order to preserve the level differences which naturally occur when speaking through some sort of face covering.

From the two simultaneous continuous HD colour video recordings, this experiment used the footage in which the talkers were facing the camera. To recall, the camera had been positioned so that the images consisted of the talker's entire head and shoulders in the centre of the screen. As the computer monitor for stimulus prompting was placed directly below the camera lens, the impression was given that the talkers were looking into the lens. The videos were cut and saved as individual files containing one stimulus sentence each (*He said [stimulus]*.). Where applicable, this was done so as to ensure that the beginning and end of each sentence showed the talker with a neutral facial expression and the mouth closed. Both video and audio data were edited and saved as AVI container files using *Canopus Edius v5.51* (25f/s, 1280x720).[34] The duration of each resultant file was 2.2s.

Two types of stimuli were produced from these recordings: auditory-only (AO) and auditory-visual (AV). The former (AO) were obtained by automatically extracting the audio streams from the corresponding videos using *FFmpeg.*[35] The high quality of the material allowed facial cues, which encode fine phonetic detail in the talker's face (e.g. lip protrusion, chin wrinkling, cheek puffing) to be clearly visible.

As the reader will recall from Chapter 3, the speech material consisted of /C_1ɑ:C_2/ nonsense syllables embedded utterance-finally in the carrier sentence *He said [stimulus]*. The consonants under investigation were /p b t d k ɡ f v s z ʃ ʒ θ ð m n/ (see Miller & Nicely, 1955). Note that /h/ and /ŋ/ were excluded from this study (necessary to constrain the length of the experiment). The target stimuli were two tokens of each of the 16 consonants produced in syllable onset position (/C_1/). Onsets were chosen so as to match the speech material examined in the acoustic study presented in Chapter 4, and because consonants are generally more easily identifiable

[34] The software is available from: http://www.goo.gl/3U61hJ [Accessed: 7th May 2014].
[35] The software is available from: http://www.goo.gl/OYtRhd [Accessed: 7th May 2014].

from onsets than from other syllable positions (Redford & Diehl, 1999; Benkí, 2003; Smits *et al.*, 2003; Weber & Smits, 2003; Woods *et al.*, 2010).

The syllable nucleus was always the open back vowel /ɑː/. This was deemed to be beneficial, because a consistent phonetic environment for consonant perception was thus ensured and coarticulatory differences were minimised (Preminger *et al.*, 1998; Fitzpatrick & Kim, 2010; Woods *et al.*, 2010). Also, the wide jaw opening for the production of this vowel enabled articulatory movements to be maximally visible (Fagel, 2005). Furthermore, the use of nonsense words (logatoms) was considered advantageous because the scope of the study was to examine the extent to which listeners rely on *acoustic* cues to consonant identity, rather than contextual and higher-level linguistic factors, such as lexical or syntactic predictability (see also §3.1.3). Lastly, impressionistic auditory judgements of the stimuli before testing made sure that no mispronunciations had occurred.

All ten talkers recorded for the AVFC corpus were included in this study. It will be recalled that all were native English speakers, their average age was 26.5 (*SD* = 5.7), all of them had had previous IPA training, and none of them reported prior experience of wearing any type of facewear on a regular basis. The two tokens per consonant were selected so that they had been produced by two different talkers. This aimed at taking into account likely variability between talkers (the speech productions of different talkers can be differently affected by facewear), and also possible idiosyncrasies. Regarding the latter, the intention was to compensate for the possibility that listeners learned the pronunciation of one talker, and for the fact that some talkers are easier to speechread than others (see also Gagné *et al.*, 1994; Kricos, 1996; Preminger *et al.*, 1998; Yakel *et al.*, 2000). To avoid bias, it was checked that all participants were unfamiliar with the talkers (see e.g. Lovitt & Allen, 2006).

All eight types of facewear included in the AVFC corpus were tested: both balaclavas (with and without a mouth hole), the motorcycle helmet, hoodie/scarf combination, *niqāb*, rubber mask, surgical mask, and the piece of tape across the mouth. The study also included the control condition (unconcealed face) in order to provide a baseline for comparison with the results from the facewear conditions.

In sum, Experiment 3 tested lay listeners' performance in consonant identification when the consonants were presented in two modalities (auditory-only, auditory-visual). Within each modality there were nine facewear conditions (control + eight types of facewear). Each facewear condition consisted of 32 items (16 consonants x 2 tokens). Hence, the test material was comprised of 576 items.

5.2.1.3 Procedure

Prior to taking part in the experiment, participants were informed about the procedure so that they could grant their informed consent to participate. Both verbal and written instructions were given, and these were formulated in such a way as to avoid biasing the participants towards one modality or the other (see e.g. Massaro, 1998; Tiippana *et al.*, 2004).

Participants were advised that the task in each trial of the forced-choice experiment was to identify only the first (onset) consonant in the test syllable (an example was given). They were instructed to click one of the response items in a 2x8 grid presented on a computer screen to choose their answer (see Figure 5.6). The response items displayed the 16 consonants in orthographic representation (<p b t d k g f v s z sh zh th dh m n>), and also embedded in example words (minimal pairs where possible, i.e., *pit/bit*, *tie/die*, *kite/guide*, *few/view*, *sip/zip*, *she/genre*, *thin/this*, *map/nap*). The example words were chosen to merely illustrate which consonant sounds the orthographic strings referred to; the use of IPA symbols was not feasible as participants lacked phonetic training. Note that participants could click either on the words or the letter buttons to make their choice. Also, the experiment was not timed. However, to help to minimise the time taken by participants to choose the desired response, the items were positioned in the grid according to their manner of articulation (plosives, fricatives, nasals) and voicing features (voiceless items were

presented on the left-hand side, and voiced items except /m/ on the right-hand side of the grid).[36]

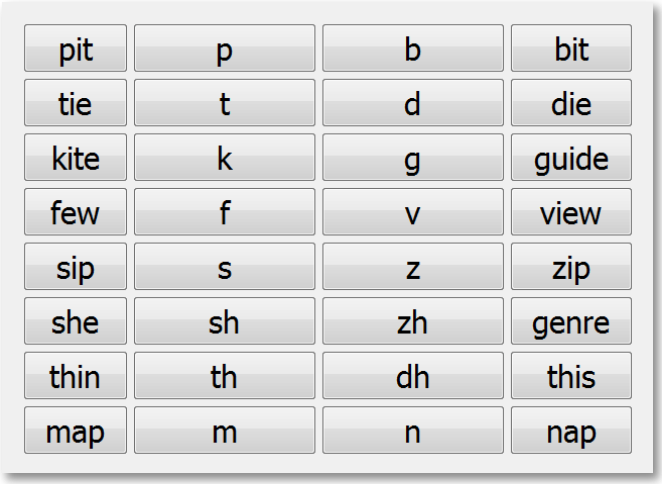

Figure 5.6. Response panel illustrating the 16 response items that were presented to participants in both forced-choice consonant identification experiments presented in this thesis (Experiments 3 and 4). In each experimental trial, participants selected their desired response by clicking one of the consonant items shown as orthographic strings (or the corresponding example words). The consonants were positioned in the grid according to their manner of articulation and voicing features.

To familiarise the participants with the experimental interface and procedure, they firstly completed a practice session (consisting of five AO and five AV control items). During the practice trials they also had the possibility to adjust the playback volume to a comfortable hearing level. The main experiment was then presented in three blocks. Between each block participants took a short rest break during which they had an informal conversation with the experimenter (the author), which merely aimed at distracting them from the task.

To compensate for practice and fatigue effects, the order of experimental trials was pseudo-randomised for each participant. No feedback about the correctness of responses was given to them. The experiment was run in a quiet computer laboratory at the Department of Language and Linguistic Science, University of York, United

[36] It was initially suspected that the ‘th’ button (representing /θ/) and the ‘dh’ button (denoting /ð/) might be confused by participants with the response buttons for ‘t’ and ‘d’. However, this was not confirmed in the later data analysis.

Kingdom. Audio was played back through Sennheiser HD 280 PRO headphones, and videos were presented on 22-inch Iiyama ProLite E2210HDS LCD monitors. The test was run using experimental control software specifically designed for the purpose of this study using the *wxLua* scripting language.[37] The entire experiment, including (de)briefing and breaks, took approximately 1.5–2hrs to complete. The study was approved by the University of York Humanities and Social Sciences Ethics Committee (for accompanying documentation see Appendices B.3 and B.4).

5.2.2 Results

The performance measure calculated to express the participants' ability to accurately identify the consonants was 'percentage correct'. The accuracy scores were analysed by conducting a series of three-way repeated-measures analyses of variance (ANOVAs) using *IBM SPSS Statistics V.19.0.0.1*, with 'modality' (AO, AV), 'facewear' (control, balaclava with and without mouth hole, helmet, hoodie/scarf, *niqāb*, rubber mask, surgical mask, tape), and 'consonant' (/p b t d k g f v s z ʃ ʒ θ ð m n/) as independent within-group factors.

Effects are reported as significant when $p < .05$. Where Mauchly's test indicated that the assumption of sphericity had been violated, the degrees of freedom, *p*-values and effect sizes (η_p^2) were adjusted using the Greenhouse-Geisser correction (the correction factor ε is listed in the corresponding results table in such cases). All results were averaged across participants. The dataset produced by one female participant (who was feeling unwell during participation) was excluded from the analysis as her results deviated significantly from the rest of the participants (statistical outliers were defined as those falling into the 1.5 interquartile ranges below the 25th and above the 75th percentile).

[37] Thanks to Tai Chi Minh Ralph Eastwood for developing the experimental control software and making it accessible for free download on *GitHub* at http://www.goo.gl/fdGsbq [Accessed: 7th May 2014].

5.2.2.1 Percentage correct (overall)

The percentage correct scores obtained in Experiment 3 were very high overall, which means that the listeners' consonant identification performance was very good. The participants reached near-ceiling performance, with 92.3% correct identification on average, 92% in the AO condition, and 92.5% in the AV condition. As such, the experiment established that the lay listeners tested here were very successful in accurately identifying spoken English consonants, irrespective of the modality these were presented in, and even when the majority of them had originally been produced through facewear.

The results of the statistical analysis of the percentage correct data obtained in Experiment 3 are shown in Appendix D.5 (see Table D.41). It was found that there was a weak but significant main effect of modality on consonant identification [$F(1,42) = 5.11$, $p < .05$, $\eta_p^2 = .11$]. This indicates that the participants on average correctly identified more consonants when they could see the talker's face, compared to when they only heard the talker's voice.

The statistical analysis furthermore revealed that the main effect of facewear [$F(6,239) = 87.43$, $p < .001$, $\eta_p^2 = .68$] was significant. This means that averaged across consonants and modalities, the various types of face coverings significantly affected the listeners' performance in the task. They accurately recognised consonants more often in some of the facewear conditions than in others. Moreover, the main effect of consonant [$F(3,120) = 26.90$, $p < .001$, $\eta_p^2 = .39$] was significant, indicating that on average some consonants were better identified than others (irrespective of facewear condition and modality).

Finally, the interactions between facewear and consonant [$F(120,5040) = 11.37$, $p < .001$, $\eta_p^2 = .21$], and between modality, facewear and consonant [$F(120,5040) = 1.23$, $p < .05$, $\eta_p^2 = .03$], were found to be significant. This suggests a complex relationship between the three variables. Specifically, the recognisability of a consonant will be dependent on the modality it has been presented in, and on the type of facewear the talker's face was occluded by when the consonant was uttered. To explore these

significant interactions in more detail, the dataset was subsequently split up by type of facewear. The results of this analysis are reported in the following section.

5.2.2.2 Percentage correct (by facewear)

The results of two-way repeated-measures ANOVAs run for the AO and AV conditions separately are listed in Appendix D.5 (see Table D.42). In the AO condition, the main effects of facewear [$F(6,245) = 56.20$, $p < .001$, $\eta_p^2 = .57$] and consonant [$F(3,134) = 27.65$, $p < .001$, $\eta_p^2 = .40$], as well as the facewear x consonant interaction [$F(120,5040) = 7.67$, $p < .001$, $\eta_p^2 = .15$], were significant. Similarly, in the AV condition, there was a significant main effect of facewear [$F(6,246) = 38.28$, $p < .001$, $\eta_p^2 = .48$] and consonant [$F(3,119) = 23.78$, $p < .001$, $\eta_p^2 = .36$], and a significant facewear x consonant interaction [$F(120,5040) = 6.27$, $p < .001$, $\eta_p^2 = .13$]. The significant main effects indicate that the effect of facewear (i.e., that some types of masks influenced consonant recognition more than others) and the effect of consonant (i.e., that some consonants were better identified than others) occurred irrespective of the modality the stimuli were presented in (AO or AV).

In subsequent *post-hoc* Bonferroni-adjusted pairwise comparisons, the results pooled by facewear type were each compared to the control condition. This test sought to establish whether the participants' performance in each of the various facewear conditions significantly differed from the baseline. It was found that in both the AO and AV conditions only the accuracy scores obtained for the tape significantly differed from the baseline (ps < .001). This implies that AO and AV consonant identification accuracy (on average) significantly decreased when the speech was produced through the tape, but that this was not the case for any of the other tested face coverings (i.e., they did not significantly affect consonant identification).

Next, two-way repeated-measures ANOVAs were run for the control and eight facewear conditions separately. The results are shown in Appendix D.5 (see Table D.43). As can be seen in the table, the effect of modality was significant only for the

tape [$F(1,42) = 6.45$, $p < .05$, $\eta_p^2 = .13$]. This implies that speech intelligibility overall improved (when visual speech cues were provided) only when the speech had originally been produced with the talker's mouth taped closed. For all other types of facewear (where performance was already very high in the AO condition) there was no statistically significant improvement – and often no improvement at all – in the AV compared to the AO condition. These findings are illustrated by the solid black (AO) and black hatched (AV) bars in Figure 5.7. They will be discussed in §5.6.1.

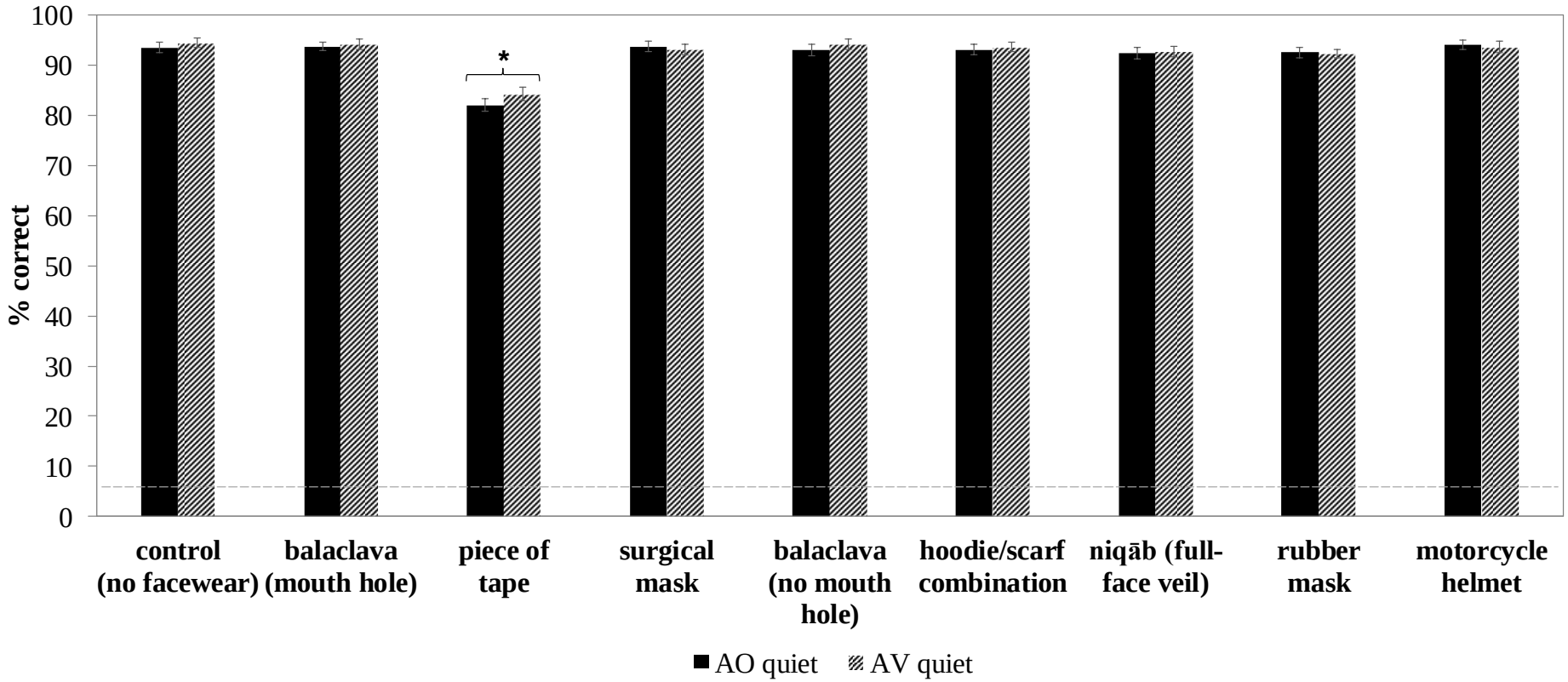

Figure 5.7. Consonant identification accuracy averaged across consonants that were presented in the quiet listening condition (Experiment 3), for each facewear condition (including control) separately, as a function of modality. The dashed horizontal line represents chance level (6%). '*' denotes a significant 'AV effect' at $p < .05$. The error bars show the standard error of the mean.

5.3 Experiment 4: Auditory-visual consonant identification in speech-in-noise conditions

The second perception experiment builds on the findings of Experiment 3 (see §5.2). Here, the same set of stimuli is tested, but this time the speech is embedded in background noise. The experiment again tests the participants' ability to identify consonants spoken through various face masks when presented in AO or AV conditions. However, the listening conditions are now considerably degraded. The goal is to determine the contribution of facial speech cues when participants have to rely to a much greater extent on the visual input from the talker's articulating face owing to the (anticipated) decrease in auditory intelligibility caused by the addition of noise. The next sections give an account of the method and results of the study.

5.3.1 Method

5.3.1.1 Participants

Forty-three native English-speaking students (35 females, 8 males) from the University of Western Sydney, Australia, participated in the experiment.[38] They were on average 19.9 years old (*SD* = 3.1) and reported normal or corrected-to-normal vision and no history of hearing impairment. None of them indicated previous experience of regularly wearing any type of facewear, or interacting with people who do so. All participants took part in the study in return for course credit. The responses of two female and two male participants had to be excluded from the final dataset owing to technical difficulties during experimentation.

[38] This work was conducted in 2012 during the author's secondment at the MARCS Institute, University of Western Sydney, Australia, as part of her contractual obligation as a member of the Marie Curie Initial Training Network 'Bayesian Biometrics for Forensics (BBfor2)'.

5.3.1.2 Speech material

The speech material was the same as that described for Experiment 3, with the exception that the audio streams in both the AO and AV conditions had background noise superimposed upon them. The speech was masked by multi-talker babble (more specifically, 8-talker babble). Babble has been shown to be a particularly effective masker, because it acts both as a powerful *informational* and an effective *energetic* masker (see also §5.6.1.2). It has been argued that compared to white or pink noise, babble reflects difficult listening conditions in a more natural way (Cutler *et al.*, 2004; Cooke, 2006; Lecumberri & Cooke, 2006).

The babble noise consisted of recordings of four females and four males speaking aloud while solving a Sudoku puzzle. The pauses were removed, and the recordings were normalised to the same RMS (root mean square) level before being mixed together. 30s of the resultant babble soundtrack was upsampled to 48kHz (from 25kHz), and a random segment was selected to be added to each stimulus file. All noise fragments had the same RMS level when mixed with the speech. The original speech stimuli were 'on average' normalised for level. This means that the RMS energies of each talker's control samples were computed based on the *He said* frames of the test sentences. The mean RMS energy levels calculated from these multiple control samples per talker were then taken as the scale factors to normalise all speech samples (including the facewear conditions) on a per-talker basis. After this, the rescaled speech was mixed with the babble using *Matlab*.[39] The mixed files were not normalised, and the noise level was kept constant. Consequently, the natural variations in the speech levels caused by the facewear were maintained during testing ($\bar{x}$ = –10.8dB SPL, *SD* = 4.8; calculated with pauses included).[40]

Finally, the visual test items were created by realigning the new 'noisy' audio streams with the original videos using *VirtualDub 1.9.11*.[41]

[39] The software is available from: http://www.goo.gl/44IBCm [Accessed: 7th May 2014].
[40] The author is very grateful to Martin Cooke for providing the babble noise soundtrack, and for offering great help with the sound mixing.
[41] The software is available from: http://www.goo.gl/ZZ4RpC [Accessed: 7th May 2014].

5.3.1.3 Procedure

The procedure was the same as described for Experiment 3. Here, participants were tested individually in a sound-attenuated IAC (Industrial Acoustics Company) booth at the MARCS Institute, University of Western Sydney, Australia. Audio was played back through Sennheiser HD 650 headphones, and videos were presented on a 22-inch BenQ E2200HD LCD monitor. The experiment was approved by the University of Western Sydney Human Research Ethics Committee (for accompanying documentation see Appendices B.5 and B.6).

5.3.2 Results

The data obtained in Experiment 4 were analysed by means of three-way repeated-measures ANOVAs following the specifications given for Experiment 3 in §5.2.2.

5.3.2.1 Percentage correct (overall)

The results of the statistical analysis of the speech-in-noise data are presented in Appendix D.5 (see Table D.44). There were again significant main effects of modality [$F(1,38) = 196.12$, $p < .001$, $\eta_p^2 = .84$], facewear [$F(5,207) = 291.93$, $p < .001$, $\eta_p^2 = .89$] and consonant [$F(10,378) = 105.96$, $p < .001$, $\eta_p^2 = .74$] on the consonant responses. The modality x facewear [$F(8,304) = 37.13$, $p < .001$, $\eta_p^2 = .49$], modality x consonant [$F(10,368) = 7.70$, $p < .001$, $\eta_p^2 = .17$], facewear x consonant [$F(120,4560) = 24.01$, $p < .001$, $\eta_p^2 = .39$], and modality x facewear x consonant [$F(120,4560) = 4.81$, $p < .001$, $\eta_p^2 = .11$] interactions were all significant.

The percentage correct scores for the AO and AV speech-in-noise data, averaged across consonants and facewear, are shown in Figure 5.8 (along with the ‘quiet’ data

for easier comparison). It becomes immediately evident from the figure that in both modalities the listeners' ability to correctly recognise consonants was greatly diminished when the stimuli were presented in noise, compared to when they were presented in the quiet listening condition. When the original soundtracks were embedded in 8-talker babble noise, the percentage correct scores substantially dropped, to 39.2% on average, 35.6% in the AO condition, and 42.7% in the AV condition. However, it seems worth pointing out that participants still performed at well above chance levels. Chance level in Experiments 3 and 4 was 6% (1/16 consonants).

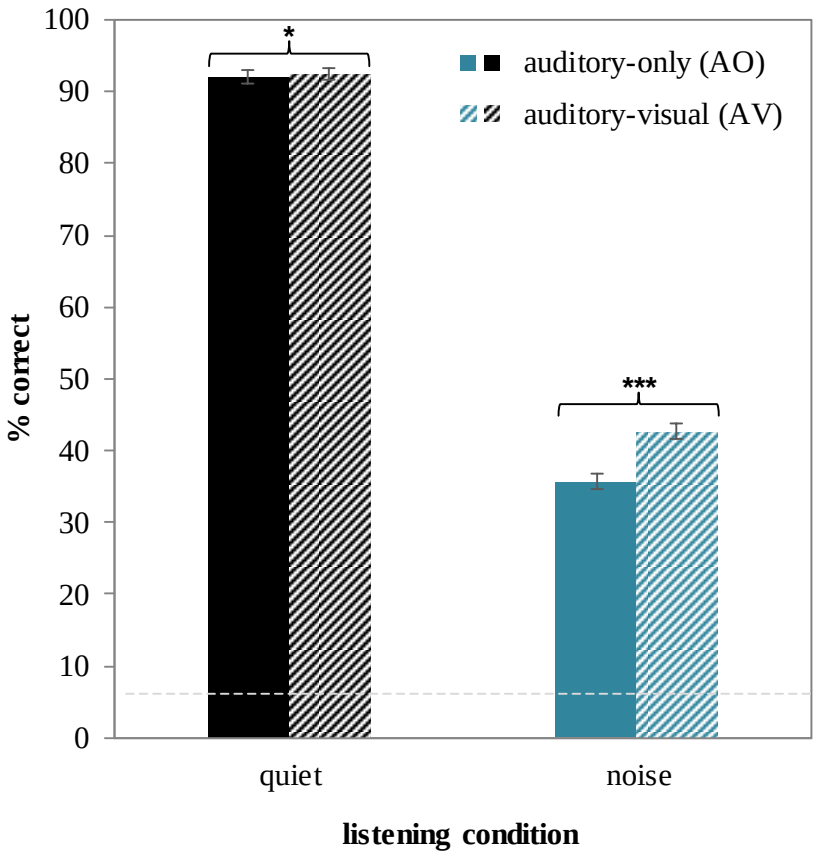

Figure 5.8. Consonant identification accuracy averaged across consonants and facewear, for each listening condition (quiet = Experiment 3, noise = Experiment 4) separately, as a function of modality. The dashed horizontal line represents chance level (6%). '***' denotes a significant 'AV effect' at $p < .001$, and '*' at $p < .05$. The error bars show the standard error of the mean.

At this stage, the crucial difference between the results obtained in Experiment 4 and those reported for Experiment 3 was – apart from the considerable overall drop in

performance – the highly significant effect of modality ($p < .001$). This indicates that the listeners (on average) identified the consonants significantly more often correctly when the talker's face was visible, compared to when they only listened to the talker's voice. This effect is shown in the two rightmost bars in Figure 5.8 (AO = solid blue, AV = blue hatched). The figure illustrates that the increase in intelligibility in the AV condition was by and large much greater in the speech-in-noise condition than in the quiet listening condition. This implies that having access to visual speech cues encoded in the talker's face, in addition to acoustic cues to consonant identity, greatly helped the listeners in overcoming the difficulties in recognising the consonants when these were presented in noise. The consonant intelligibility gain in the AV condition is from now on referred to as the 'AV effect'.

5.3.2.2 Percentage correct (by facewear)

Once again, ANOVAs were rerun for the AO and AV conditions separately. The results of these tests are shown in Appendix D.5 (see Table D.45). In the AO condition, the main effects of facewear [$F(7,213) = 145.85$, $p < .001$, $\eta_p^2 = .79$] and consonant [$F(15,570) = 80.30$, $p < .001$, $\eta_p^2 = .68$] were significant, as was the interaction between facewear and modality [$F(120,4560) = 14.65$, $p < .001$, $\eta_p^2 = .28$]. Likewise, in the AV condition, there was a significant main effect of facewear [$F(8,304) = 262.86$, $p < .001$, $\eta_p^2 = .87$] and consonant [$F(10,370) = 94.68$, $p < .001$, $\eta_p^2 = .71$], and a significant facewear x modality interaction [$F(120,4560) = 18.06$, $p < .001$, $\eta_p^2 = .32$].

Post-hoc comparisons (Bonferroni) revealed that in the AV condition, the accuracy scores for all types of facewear were significantly lower than in the control condition (ps < .001). This suggests that the impoverished visual speech cues induced by all types of facial occlusions had an adverse effect on consonant identification in noise. In the AO data, on the other hand, the recognition rates obtained in only some of the facewear conditions were significantly lower than the rates obtained in the baseline. These were the tape, rubber mask, helmet (ps < .001), *niqāb* ($p < .01$), and the

balaclava with the mouth hole ($p < .05$) conditions. This implies that the changes to the acoustic properties of the consonants (caused by acoustic absorption and/or modified speech production) disturbed consonant identification only when the consonants had been produced through these four face coverings.

ANOVAs were also carried out for the control and all facewear conditions individually in order to determine whether consonant identification accuracy in the AV condition increased for any of the conditions. The results are shown in Appendix D.5 (see Table D.46). It was found that the main effect of modality on consonant identification was significant in the control [$F(1,38) = 146.09$, $p < .001$, $\eta_p^2 = .79$], tape [$F(1,38) = 134.77$, $p < .001$, $\eta_p^2 = .78$], and balaclava (mouth hole) conditions [$F(1,38) = 130.64$, $p < .001$, $\eta_p^2 = .78$]. This finding is illustrated in Figure 5.9, where the solid blue bars denote the AO and the blue hatched bars the AV condition. As can be seen in the figure, the bar for the AV condition is in all three facewear conditions considerably higher than the bar for the corresponding AO condition. This signifies a significant improvement in consonant identification from the AO to the AV modality, and thus affirms an especially strong AV effect for the control, tape and balaclava (mouth hole) conditions.

Furthermore, statistical analysis revealed a significant difference between AO and AV consonant identification when the speech was produced through the balaclava (no mouth hole) [$F(1,38) = 7.80$, $p < .01$, $\eta_p^2 = .17$], surgical mask [$F(1,38) = 8.12$, $p < .01$, $\eta_p^2 = .18$], and hoodie/scarf [$F(1,38) = 6.33$, $p < .05$, $\eta_p^2 = .14$]. This means that in these conditions consonant recognition was again significantly improved when the talker's disguised face was presented, i.e., participants' performance increased from AO to AV. However, the AV effect was overall less pronounced than for the control, balaclava (mouth hole) and tape conditions presented earlier.

Finally, no consonant intelligibility gain when the face was presented (AV effect) was observed for speech produced through the helmet ($p = .762$), rubber mask ($p = .536$), and *niqāb* ($p = .488$). Here, it made no difference whether the participants only listened to the talker's voice, or additionally saw the talker's face. Their performance in each of these three facewear conditions was equally low in the AO and AV conditions (see Figure 5.9). These results will be further discussed in §5.6.1.2.

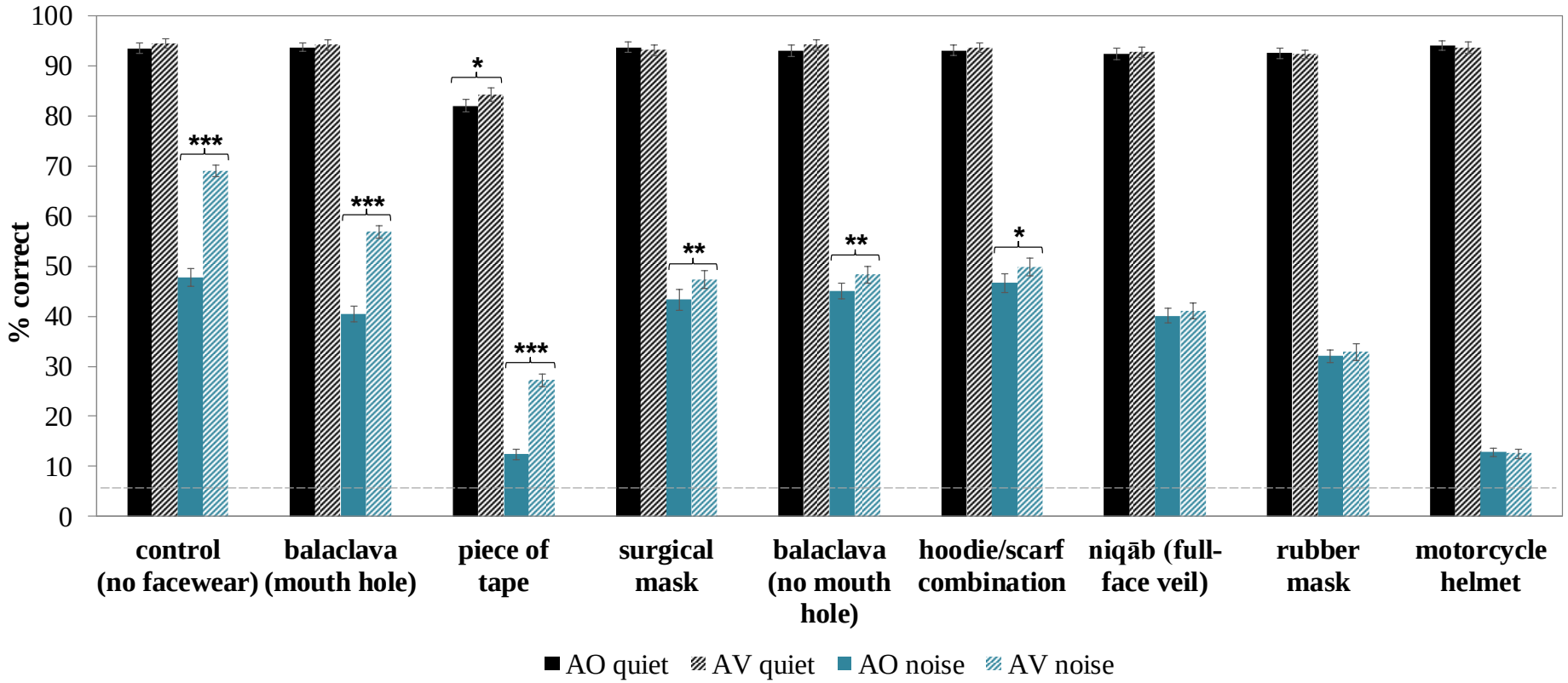

Figure 5.9. Consonant identification accuracy averaged across consonants, for each listening condition (quiet = Experiment 3, noise = Experiment 4) and facewear condition (including control) separately, as a function of modality. The dashed horizontal line represents chance level (6%). '***' denotes a significant 'AV effect' at $p < .001$, '**' at $p < .01$, and '*' at $p < .05$. The error bars show the standard error of the mean.

5.4 Consonant identification performance

To observe how accurately listeners can identify individual consonants and vowels is one of the traditional methodologies applied in the study of auditory speech perception. In recent years, this method has also been widely adopted in research on auditory-visual speech processing. Empirical research on perceptual errors that occur while human listeners perceive consonants and vowels dates back to Miller and Nicely's frequently-cited 1955 study. The main outcome of this work was that the identifiability of consonants presented in noise varies substantially, and that listeners mishear the sounds in systematic and predictable ways. Miller and Nicely presented 16 consonants embedded in /Cɑː/ syllables to listeners and asked them what consonants they perceived. The stimuli in their study were masked with wideband noise at signal-to-noise ratios (SNRs) ranging from –18dB to +12dB. The researchers found, firstly, that some consonants (e.g. non-sibilants) were difficult to identify even at high SNRs, while others (e.g. sibilants and nasals) were accurately identified even at much lower SNRs. Secondly, as the SNRs decreased, participants only chose from a subset of possible consonant responses.

The basic findings from Miller and Nicely's classic experiment were subsequently affirmed by a vast range of studies on perceptual phoneme confusions. The parameters which were often experimentally manipulated include different syllable positions (Redford & Diehl, 1999; Benkí, 2003; Smits *et al.*, 2003; Weber & Smits, 2003; Woods *et al.*, 2010), the listener's native language (Cutler *et al.*, 2007, 2008; Fitzpatrick & Kim, 2010), and most commonly, the type and level of background noise that is masking the speech (Wang & Bilger, 1973; Soli & Arabie, 1979; Dubno & Levitt, 1981; Benkí, 2003; Weber & Smits, 2003; Cutler *et al.*, 2004; Simpson & Cooke, 2005; Lecumberri & Cooke, 2006; Lovitt & Allen, 2006; Cutler *et al.*, 2007; Phatak & Allen, 2007; Cutler *et al.*, 2008; Phatak *et al.*, 2008; Kim *et al.*, 2009; Fitzpatrick & Kim, 2010; Peláez-Moreno *et al.*, 2010).

So far, the results of the two consonant identification experiments presented in this chapter were in both cases averaged across the 16 tested consonants. Looking in more detail at the total number of consonant identification errors that participants made, one finds that among the 24,768 responses returned by 43 participants in the

quiet listening condition, a total of 1,931 errors occurred. By contrast, of the 22,464 consonant responses provided by 39 participants in the speech-in-noise test, 13,665 were erroneous. In the AO conditions, respectively, participants made a total of 993 errors when the listening conditions were good, and 7,234 when the conditions were degraded by noise. In the AV conditions, participants falsely identified 938 consonants in the quiet and 6,431 consonants in the noise condition.

The aim of the following sections is to give a full description of the data obtained in Experiments 3 and 4 on the consonant level. Large datasets of the kind created in this study call for powerful tools to analyse the underlying patterns of, in this case, consonant identification errors. Common procedures are hence introduced. Both quantitative and qualitative accounts of the most common perceptual errors are given, and the effects of facewear on consonant identification are discussed.

5.4.1 Percentage correct (per consonant)

To examine the consonant identification results in more depth, the data from each experiment were, as a first step, broken down by consonant. The outcome is shown in Figure 5.10. The figure reveals the percentage correct scores averaged across facewear as a function of listening condition (quiet/noise) and modality (AO/AV) for each of the 16 consonants separately. The consonants are ordered along the x-axis according to the mean for AO+AV per consonant in the quiet listening condition (in descending order).

Figure 5.10 shows that the patterns were much more heterogeneous in the noise (blue) than in the quiet (black) condition. Overall, consonant intelligibility was enhanced when the talker's face was visible to participants (AV), relative to when they only heard the talker's voice (AO). This trend emerged only for some of the consonants in the quiet condition, in particular the fricatives /θ/ and /ð/. In the speech-in-noise data, the AV effect was much more evident. Here, most of the

consonants were notably better recognised when participants had access to visual speech cues from the talker's articulating face.

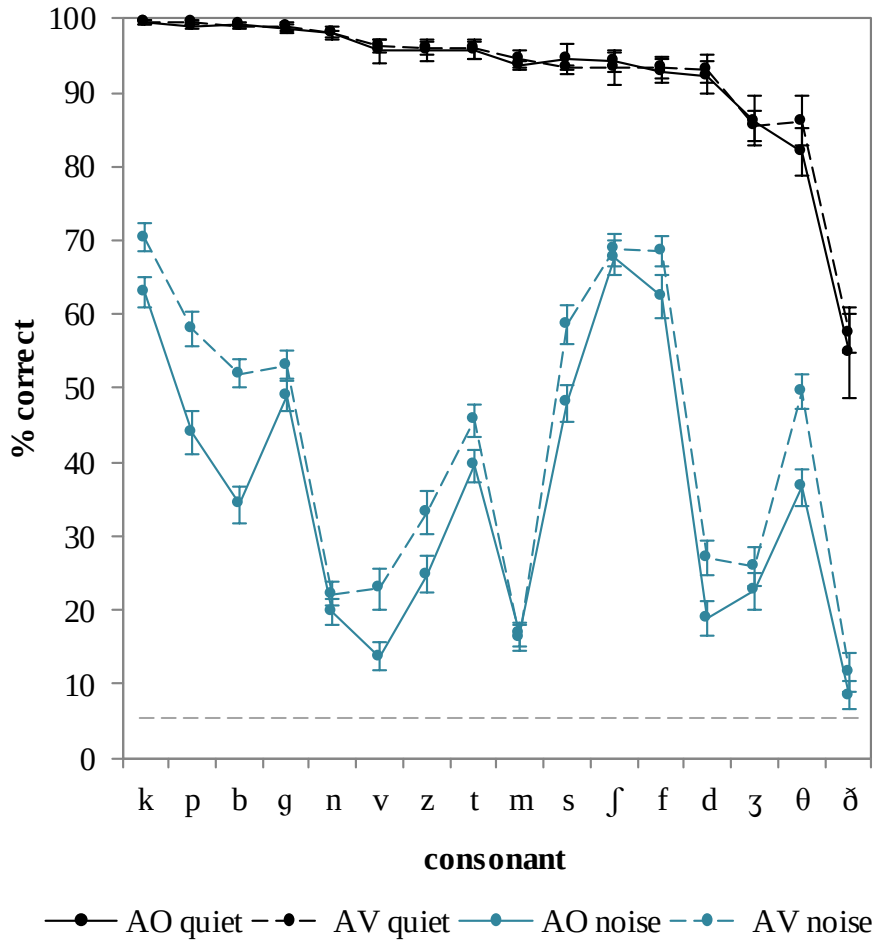

Figure 5.10. Consonant identification accuracy averaged across facewear, for each consonant separately, as a function of listening condition (quiet = black, noise = blue) and modality (AO = solid lines, AV = hatched lines). The consonants are ordered along the x-axis according to the mean for AO + AV per consonant in the quiet listening condition (in descending order). The dashed horizontal line represents chance level (6%). The error bars show the standard error of the mean.

To be in a position to evaluate the specific nature of the consonant identification errors across facewear conditions – and hence the participants' performance in the task – in a systematic and insightful way, the data were subsequently reorganised in *confusion matrices*. These are presented and discussed in the next section.

5.4.2 Confusion matrices

Consonant confusion matrices were constructed in *R*[42] for each listening condition, modality and facewear condition (+ control) separately. This resulted in a total of 36 tables (2 listening conditions x 2 modalities x 1+8 facewear conditions). The matrices containing the data produced in the control (no facewear) condition are shown in Tables 5.1 to 5.4. Owing to the large number of tables, all other tables (including larger versions of Tables 5.1 to 5.4) can be consulted in Appendix D.1.

The benefit of arranging consonant responses in the form of confusion matrices is that they illustrate both how often a consonant was correctly identified, and how often the same consonant was falsely perceived as another consonant. In each matrix, the tested consonants are displayed in rows, and participants' responses in columns (same consonant order). Consequently, the correct responses are shown along the diagonal of each matrix, and the incorrect responses on either side of the diagonal. The number in each cell is the frequency with which a particular stimulus-response pair occurred (total counts).

[42] Available from: http://www.goo.gl/pFxVgK [Accessed: 7th May 2014].

	response							**control, quiet, AO**								
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	82	3	0	0	0	0	0	0	0	0	0	1	0	0	0	86
ð	0	0	47	0	0	0	0	0	0	0	0	2	28	8	0	1	86
f	0	0	0	83	0	0	0	0	0	0	0	0	3	0	0	0	86
g	0	0	0	0	85	0	0	0	0	0	0	0	0	0	0	1	86
k	0	0	0	0	0	85	0	0	0	0	0	0	1	0	0	0	86
m	0	0	0	0	0	0	85	0	1	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	81	2	2	0	0	1	0	86
ʃ	0	0	0	0	0	0	0	0	0	1	85	0	0	0	0	0	86
t	0	0	0	0	0	0	0	0	0	0	0	83	3	0	0	0	86
θ	0	0	1	9	0	0	0	0	0	0	0	2	73	1	0	0	86
v	0	0	1	0	1	0	0	0	0	0	0	0	0	84	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	83	3	86
ʒ	0	0	1	0	7	0	0	0	0	0	0	0	0	0	0	78	86
total	86	82	53	92	93	85	85	86	87	82	87	89	109	93	84	83	1376

Table 5.1. Confusion matrix for the consonants presented auditorily in the control condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant). Note that a larger version of this table can be found in Appendix D.1 (Table D.3).

	response							**control, quiet, AV**								
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	81	4	0	1	0	0	0	0	0	0	0	0	0	0	0	86
ð	0	0	52	0	0	0	0	0	0	0	0	0	33	1	0	0	86
f	0	0	0	82	0	0	0	0	0	0	0	0	2	2	0	0	86
g	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	85	0	0	0	0	1	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	85	0	0	1	0	0	0	0	86
s	0	0	1	0	0	0	0	0	0	81	2	0	0	0	2	0	86
ʃ	0	0	0	0	0	0	0	0	0	4	80	0	0	0	0	2	86
t	0	0	0	0	0	0	0	0	0	0	0	85	1	0	0	0	86
θ	0	0	2	1	0	0	0	0	0	0	0	1	82	0	0	0	86
v	0	0	0	1	0	0	0	0	0	0	0	0	0	85	0	0	86
z	0	0	0	0	0	0	0	0	0	0	1	0	0	0	82	3	86
ʒ	0	0	1	0	5	0	0	0	0	0	0	0	0	0	2	78	86
total	86	81	60	84	92	85	86	86	85	85	84	87	118	88	86	83	1376

Table 5.2. Confusion matrix for the consonants presented auditory-visually in the control condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant). Note that a larger version of this table can be found in Appendix D.1 (Table D.12).

	response								**control, noise, AO**								
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	43	2	4	2	3	3	0	0	0	7	3	1	1	3	2	4	78
d	3	14	6	1	15	1	0	1	1	14	3	0	13	1	2	3	78
ð	4	6	14	4	4	2	1	1	2	1	1	1	17	14	6	0	78
f	2	0	1	62	0	0	0	2	0	1	0	0	4	6	0	0	78
g	1	3	0	1	64	1	0	0	0	5	1	0	0	0	1	1	78
k	0	0	0	2	0	74	0	0	1	1	0	0	0	0	0	0	78
m	23	2	1	5	5	6	16	1	6	2	1	2	4	3	1	0	78
n	6	9	2	3	3	7	3	20	5	2	0	4	8	5	0	1	78
p	1	1	1	2	4	23	1	0	37	1	1	4	1	1	0	0	78
s	0	0	1	5	0	0	0	2	0	57	2	0	5	0	4	2	78
ʃ	0	4	0	1	0	0	0	0	0	1	67	0	2	1	0	2	78
t	1	0	1	2	0	19	0	0	16	0	0	36	2	1	0	0	78
θ	0	1	2	26	1	0	0	0	0	3	3	1	36	4	1	0	78
v	7	2	2	34	0	1	1	0	1	0	0	0	9	21	0	0	78
z	1	5	7	4	0	2	0	1	0	1	1	1	18	3	27	7	78
ʒ	2	10	7	2	11	9	1	5	2	3	9	1	6	1	1	8	78
total	94	59	49	156	110	148	23	33	71	99	92	51	126	64	45	28	1248

Table 5.3. Confusion matrix for the consonants presented auditorily in the control condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant). Note that a larger version of this table can be found in Appendix D.1 (Table D.21).

	response								**control, noise, AV**								
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78
d	0	35	2	0	4	0	0	1	0	12	10	0	8	0	5	1	78
ð	0	0	17	0	0	0	0	0	0	0	0	1	59	0	0	1	78
f	0	0	0	75	0	0	0	0	0	0	0	0	2	1	0	0	78
g	0	2	0	0	75	0	0	0	0	0	1	0	0	0	0	0	78
k	0	0	0	0	0	77	0	0	0	0	0	0	0	0	1	0	78
m	51	1	0	0	0	0	16	0	8	1	0	0	0	1	0	0	78
n	0	16	1	0	8	6	0	28	0	0	1	15	1	0	0	2	78
p	2	0	0	0	0	0	0	0	76	0	0	0	0	0	0	0	78
s	0	0	0	0	0	0	0	0	0	70	2	0	1	0	4	1	78
ʃ	0	0	1	0	0	0	0	0	0	1	73	0	0	0	0	3	78
t	0	1	0	0	0	14	0	0	1	0	0	56	6	0	0	0	78
θ	0	0	1	1	0	0	0	0	0	0	0	1	75	0	0	0	78
v	0	0	0	49	0	0	0	0	0	1	0	0	0	27	1	0	78
z	0	0	2	0	0	0	0	1	0	1	0	1	3	0	55	15	78
ʒ	0	7	2	2	11	2	0	1	0	4	19	0	1	0	1	28	78
total	131	62	26	127	98	99	16	31	85	90	106	74	156	29	67	51	1248

Table 5.4. Confusion matrix for the consonants presented auditory-visually in the control condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant). Note that a larger version of this table can be found in Appendix D.1 (Table D.30).

Next, all stimulus-response pairs which occurred in ≥10% of the trials in which the stimulus was presented were extracted from the confusion matrices (incorrect answers only). This equalled a minimum of 9 (out of 86 possible) instances in the quiet condition (86 = 43 participants x 2 tokens per consonant), and a minimum of 8 (out of 78 possible) instances in the noise condition (78 = 39 participants x 2 tokens per consonant). For example, in 28 of the 86 presentations of /ð/ in the control condition (quiet/AO), /ð/ was incorrectly classified as /θ/. Hence, the rate of occurrence of this type of confusion of /ð/ with another consonant was 33% (in this particular experimental condition). In sum, this analysis revealed a low count of pairs (with an occurrence rate of ≥10%) in the quiet condition (N_{AO} = 22, N_{AV} = 23), but a high count in the noise condition (N_{AO} = 247, N_{AV} = 226) across facewear conditions.

The results of this analysis are shown in Table 5.5 (quiet/AO), Table 5.6 (quiet/AV), Table 5.7 (noise/AO), and Table 5.8 (noise/AV). To emphasise once more, the percentages shown in these tables are *not* the overall identification or error rates for the target consonant (labelled 'stim'). They indicate how often a particular type of confusion occurred, i.e., how often the target consonant was misperceived as another consonant (labelled 'resp').

As can be seen in Table 5.5 (quiet/AO) and Table 5.6 (quiet/AV), the most common error in the quiet listening condition was for /ð/ to be misperceived as /θ/. In the AO condition, this error occurred in 33% of all trials where /ð/ was presented in the control condition, and to a similar extent across facewear conditions (27%, tape; 28%, hoodie/scarf; 28%, surgical mask; 29%, helmet; 31%, balaclava without mouth hole; 37%, *niqāb*; 40%, balaclava with mouth hole; 21%, rubber mask). This pattern was consistent across the corresponding AV samples. Among the facewear conditions (both in AO + AV), /θ/ was misclassified as /ð/ in only 1–8% of cases. This demonstrates the asymmetrical nature of the confusion patterns observed in the present data. That is, perceptual confusions typically occurred in one direction only (for further discussion of this finding see §5.4.3).

Similarly, /θ/ was very frequently perceived as /f/ in the control and most facewear samples (except helmet and balaclava with mouth hole in the AO and AV modalities, and control and surgical mask in the AV condition). Again, the occurrence rate for

the opposite direction of confusion (i.e., /f/ misclassified as /θ/) was always below 10%. Additionally, the speech produced through the tape was highly prone to perceptual errors (as common sense would predict). There was a particularly high rate of misperceptions of /m/, which was frequently mistaken for /v/ (43%, AO; 40%, AV). The fricative /v/, however, was for the most part correctly identified. A similar one-sided confusion occurred for /ʒ/, which was commonly confused with /z/ in the tape condition (35%, AO; 37%, AV).

Table 5.7 (noise/AO) and Table 5.8 (noise/AV) illustrate that the number of incorrect stimulus-response pairs with an occurrence rate of ≥10% increased greatly in the speech-in-noise data. By and large, most perceptual errors again occurred among the fricatives. The tables also reveal a fair amount of place of articulation and voicing errors among plosives, as well as manner and place of articulation errors between plosives and nasals. Interestingly, most errors can now be found in the helmet condition. The high number of confusions makes it difficult to extract any coherent patterns from the speech-in-noise data. The reader is therefore referred to the *d-prime* analysis presented in §5.5, which will give a more detailed account of the consonant confusions observed in the speech-in-noise data.

quiet listening condition, auditory-only															
control			**balaclava 2**			**tape**			**surgical mask**			**balaclava 1**			
stim	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	
ð	θ	33	ð	θ	40	m	v	43	ð	θ	28	ð	θ	31	
θ	f	10				ʒ	z	35	θ	f	10	θ	f	12	
						ð	θ	27							
						ʃ	s	19							
						f	θ	14							
						θ	f	12							
						ð	v	10							
			hoodie/scarf			***niqāb***			**rubber mask**			**helmet**			
			ð	θ	28	ð	θ	37	ð	θ	41	ð	θ	29	
			θ	f	16	θ	f	17	θ	f	12	ʒ	g	10	

Table 5.5. Most frequent consonant confusions in the quiet listening condition when the consonants were presented auditorily in the control and facewear conditions. The table shows all incorrect stimulus-response pairs which occurred in ≥10% of the trials in which a stimulus was presented. Note that the percentages listed in the table are not the overall identification or error rates for a particular target consonant (‘stim’). They indicate how often a particular type of confusion occurred, i.e., how often the target consonant was misperceived as another consonant (‘resp’).

quiet listening condition, auditory-visual															
control			**balaclava 2**			**tape**			**surgical mask**			**balaclava 1**			
stim	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	
ð	θ	38	ð	θ	35	m	v	40	ð	θ	29	ð	θ	31	
						ʒ	z	37				θ	f	12	
						ð	θ	29							
						ʃ	s	22							
						θ	f	14							
						d	g	13							
						f	t	12							
			hoodie/scarf			***niqāb***			**rubber mask**			**helmet**			
			ð	θ	27	ð	θ	36	ð	θ	36	ð	θ	30	
			ð	v	16	ð	v	16	θ	f	14				
			θ	f	12	θ	f	14	f	θ	10				
						f	θ	12							

Table 5.6. Most frequent consonant confusions in the quiet listening condition when the consonants were presented auditory-visually in the control and facewear conditions. The table shows all incorrect stimulus-response pairs which occurred in ≥10% of the trials in which a stimulus was presented. Note that the percentages listed in the table are not the overall identification or error rates for a particular target consonant (‘stim’). They indicate how often a particular type of confusion occurred, i.e., how often the target consonant was misperceived as another consonant (‘resp’).

speech-in-noise, auditory-only															
control			balaclava 2			tape			surgical mask			balaclava 1			
stim	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*
v	f	44	g	k	28	t	k	33	θ	f	45	θ	f	78
θ	f	33	b	f	23	g	k	29	ð	v	31	z	s	50
m	b	29	ð	f	23	s	θ	29	g	n	28	v	f	38
p	k	29	v	b	19	t	f	29	b	f	23	m	k	33
t	k	24	v	f	18	θ	f	28	ð	θ	22	s	θ	32
z	θ	23	ʒ	g	18	b	g	26	f	θ	22	ð	θ	27
ð	θ	22	t	k	17	n	f	24	v	f	18	d	θ	22
d	g	19	θ	f	17	f	θ	23	d	θ	17	ʒ	g	22
d	s	18	v	s	17	p	f	19	n	b	17	p	k	19
ð	v	18	ʒ	ʃ	17	p	k	19	v	s	17	t	k	17
d	θ	17	b	v	15	z	f	19	d	g	15	v	θ	17
ʒ	d	13	m	b	15	ð	s	18	m	p	15	b	k	14
n	d	12	m	f	15	k	p	18	n	g	14	d	g	14
v	θ	12	m	p	15	d	g	17	z	b	14	d	n	13
ʒ	k	12	g	s	14	d	s	15	n	d	13	d	b	12
ʒ	ʃ	12	s	θ	14	ð	k	15	s	z	13	n	d	12
n	θ	10	d	ð	13	ʃ	θ	15	z	θ	13	n	k	12
			d	θ	12	v	g	15	ʒ	f	13	z	ʃ	12
			g	z	12	k	f	14	b	g	12	d	k	10
			θ	z	12	m	s	14	b	k	12	ð	g	10
			b	θ	10	m	f	13	ʒ	g	12	v	b	10
			f	θ	10	m	g	13	ð	f	10	ʒ	ʃ	10
						θ	s	13	k	p	10			
						ʒ	t	13	m	b	10			
						ð	f	12	p	k	10			
						ð	ʃ	12	t	θ	10			
						g	p	12	v	ʃ	10			
						d	ʃ	10	z	f	10			
						m	k	10						
						n	b	10						
						n	s	10						
						ʃ	ʒ	10						
						v	s	10						
						ʒ	g	10						
						ʒ	k	10						

(table continues on next page)

speech-in-noise, auditory-only *(cont.)*												
hoodie/scarf			***niqāb***			**rubber mask**			**helmet**		
stim	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*
θ	f	33	θ	f	38	p	k	45	ʃ	s	29
ð	θ	32	m	k	27	θ	f	33	θ	s	22
p	k	24	t	p	23	m	f	32	f	θ	21
ð	v	22	v	f	20	d	s	27	ʒ	f	21
b	g	18	d	k	18	ʒ	g	22	d	f	19
n	k	17	ʃ	ʒ	15	t	p	18	t	k	19
s	z	17	t	k	15	ʒ	d	15	t	p	19
v	s	17	f	θ	14	d	ʃ	14	ð	f	18
d	g	15	d	g	13	g	k	14	b	f	17
m	b	15	n	g	13	n	k	14	v	b	17
m	k	15	n	k	13	ʃ	d	14	v	f	17
m	p	15	v	k	13	ʃ	ʒ	14	s	g	15
t	k	15	v	θ	13	t	f	14	v	s	15
v	t	14	ʒ	ʃ	13	t	θ	14	g	f	14
m	f	13	d	s	12	v	s	14	g	s	14
b	k	12	d	t	12	n	θ	13	θ	f	14
d	k	12	ð	s	12	b	s	12	z	g	14
ð	k	12	z	k	12	ð	θ	12	ʒ	s	14
f	θ	12	z	θ	12	g	d	12	m	k	13
n	g	12	z	v	12	n	p	12	n	g	13
v	f	12	ʒ	g	12	d	θ	10	n	s	13
v	z	12	ð	θ	10	ð	s	10	p	k	13
v	ʒ	12	n	d	10	ð	z	10	p	t	13
ʒ	g	12	s	θ	10	k	p	10	s	k	13
b	m	10				n	g	10	v	ʃ	13
d	n	10				n	t	10	z	k	13
m	θ	10				z	ʒ	10	ʒ	g	13
ʃ	ʒ	10				ʒ	k	10	ð	d	12
ʒ	ʃ	10				ʒ	θ	10	k	θ	12
									t	θ	12
									b	g	10
									d	f	10
									ð	k	10
									g	k	10
									k	s	10
									n	k	10
									s	d	10
									ʃ	θ	10
									t	b	10
									z	f	10
									z	m	10

Table 5.7. Most frequent consonant confusions in the speech-in-noise condition when the consonants were presented auditorily in the control and facewear conditions.

speech-in-noise, auditory-visual															
control			**balaclava 2**			**tape**			**surgical mask**			**balaclava 1**		
stim	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*
ð	θ	76	ð	θ	56	g	k	55	ð	θ	43	θ	f	74
v	f	63	m	b	49	t	k	46	b	f	31	z	s	45
m	b	56	v	f	45	v	f	36	g	n	31	v	f	36
ʒ	ʃ	24	g	k	33	ð	k	35	θ	f	28	s	θ	33
n	d	21	ʒ	ʃ	31	m	b	35	m	p	26	p	k	27
n	t	19	m	p	28	n	f	27	ʒ	g	23	ð	θ	26
z	ʒ	19	g	s	17	ʒ	t	26	ð	v	21	m	k	24
t	k	18	n	d	17	b	p	24	v	f	21	d	θ	21
d	s	15	t	k	15	ʃ	s	21	n	b	19	ð	g	21
ʒ	g	14	ʒ	g	13	θ	k	21	z	f	18	v	θ	21
d	ʃ	13	s	z	12	ð	g	19	d	g	15	d	g	17
d	θ	10	ʒ	d	12	d	ʃ	18	d	θ	14	t	k	15
m	p	10	d	θ	10	t	f	18	t	θ	13	ʒ	g	14
n	g	10	f	v	10	ʒ	g	18	n	m	12	d	b	13
			g	t	10	p	b	17	f	θ	10	n	d	13
			θ	s	10	m	f	15	f	v	10	ð	v	12
						m	v	14	k	p	10	ð	z	12
						n	k	14	m	b	10	n	k	12
						s	z	14	n	f	10	s	f	12
						d	f	13	n	θ	10	ð	d	10
						z	v	13	p	t	10	v	b	10
						d	g	12	z	θ	10			
						d	s	12						
						m	p	12						
						n	ð	12						
						p	f	12						
						d	t	10						
						n	v	10						
						ʃ	g	10						
						t	d	10						
						z	n	10						

(table continues on next page)

speech-in-noise, auditory-visual *(cont.)*											
helmet			***niqāb***			**hoodie/scarf**			**rubber mask**		
stim	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*	*stim*	*resp*	*%*
ʃ	s	36	θ	f	59	θ	f	40	b	θ	40
d	s	24	d	k	26	ð	θ	36	m	θ	38
θ	s	23	t	p	24	n	k	22	n	θ	35
n	g	19	v	f	21	p	k	22	θ	f	33
b	p	17	m	k	17	b	g	19	p	k	32
f	θ	17	m	b	15	m	b	19	g	k	31
p	k	17	n	g	15	ð	v	17	d	s	26
t	k	17	t	k	15	m	f	17	f	θ	22
z	k	17	ð	s	13	m	k	17	ʒ	g	22
ʒ	s	17	f	θ	13	ʒ	g	15	v	θ	21
t	p	16	ʃ	ʒ	13	d	g	14	ʒ	d	18
t	θ	15	v	s	13	f	θ	14	ð	g	17
θ	f	15	ʒ	g	13	b	d	13	ð	θ	17
v	ʃ	15	d	f	12	t	k	13	t	θ	17
ð	f	14	ð	θ	12	v	p	13	t	k	15
ð	g	14	ð	z	12	v	t	13	z	ʒ	15
g	d	14	s	θ	12	d	θ	12	d	ʃ	13
g	f	14	z	θ	12	m	p	12	d	θ	13
s	k	14	d	s	10	v	k	12	n	k	13
z	g	14	m	v	10	v	z	12	ʃ	ʒ	13
ʒ	k	14	n	k	10	ʒ	ʃ	12	b	s	12
ʒ	θ	14	n	s	10	b	k	10	p	θ	12
ð	t	13	p	θ	10	d	s	10	ʃ	d	12
k	g	13	v	θ	10	f	s	10	ʒ	θ	12
m	f	13	z	ð	10	v	s	10	b	ð	10
v	θ	13				z	ʒ	10	ð	d	10
ʒ	d	13							k	p	10
ʒ	f	13							n	d	10
d	θ	12							v	b	10
k	θ	12									
p	b	12									
z	d	12									
b	f	10									
f	g	10									
f	k	10									
m	d	10									
m	g	10									
m	k	10									
s	g	10									
t	d	10									
v	f	10									
v	s	10									

Table 5.8. Most frequent consonant confusions in the speech-in-noise condition when the consonants were presented auditory-visually in the control and facewear conditions.

5.4.3 Response bias

The most common perceptual error recorded in the consonant identification study was for fricatives to be confused with each other. Furthermore, some consonants were systematically identified as other consonants, i.e., certain responses to a particular target consonant were consistently favoured over others. The data revealed that considerable *asymmetries* exist among the confusable consonants. That is, in cases where consonant A was frequently misclassified as consonant B, it was not necessarily the case that B was equally often (or at all) misperceived as A. These findings accord with Miller & Nicely (1955) and other studies listed in §5.4.

Altogether, the confusion patterns found for speech produced through facewear were, as expected, highly variable. By trend, some consonants would elicit more incorrect responses (*false alarms*) than correct responses (*hits*). For example, the hit rate of 46% for /θ/ in control (noise/AO) was much higher than for /ð/ (18%). The majority of /θ/ responses, however, were elicited by consonants other than /θ/, especially /z/ (23%), /ð/ (22%), and /d/ (17%). As can be seen, /ð/ even elicited more false /θ/ responses (22%) than /ð/ hits (18%). To name another example, /v/ elicited twice as many false /f/ responses (36%) as /v/ hits (18%) in the balaclava without mouth hole condition (noise/AV), while /f/ was misclassified as /v/ in only 4% of the cases (79% correct responses for /f/).

The complex, asymmetrical response patterns that emerged in the present data reflect a considerable *response bias* on the part of the observers. Using percentage correct scores alone (as shown in Figure 5.10) is by virtue of this bias not sufficient to adequately represent and compare consonant identifiability across conditions. Presenting perceptual confusions in form of hit rates would misrepresent consonant identification accuracy, and the observers' speech perception performance in general.

In order to help to overcome the response bias, various techniques have been proposed in the literature. These include bias measures such as *d-prime*, *beta*, and *criterion* provided by signal detection theory (Benkí, 2003; Lidestam & Beskow, 2006; Woods *et al.*, 2010), sequential information analysis (Miller & Nicely, 1955; Wang & Bilger, 1973; Bernstein *et al.*, 2000; Smits, 2000; Benkí, 2003; Cutler *et al.*,

2004; Lovitt & Allen, 2006; Kim *et al.*, 2009; Fitzpatrick & Kim, 2010), analysis of the patterns of feature-processing errors in terms of single-feature versus combined place, manner, and voicing errors (Woods *et al.*, 2010), multidimensional scaling (Smits, 2000), formal concept analysis (Peláez-Moreno *et al.*, 2010), and hierarchical cluster analysis (Lidestam & Beskow, 2006).

Here, two of the most commonly applied procedures were chosen to analyse the consonant confusions in more depth. These are the information theoretical approach SINFA (Sequential INFormation Analysis), and the signal detection measure *d-prime*. Both techniques take the same basic data as input, and incorporate information from correct responses (hits) and incorrect responses (false alarms). However, they differ in terms of the calculation and interpretation of the results. In the context of this thesis, only the results of the *d-prime* analysis are discussed in the following section (mainly for reasons of lack of space).[43]

[43] The interested reader is referred to Wang & Bilger (1973) for a thorough explanation of the underlying methodology of the SINFA method. The author is very grateful to David van Leeuwen for interesting and helpful discussions about the data, and especially for providing the *R* script to run SINFA. Please note that the script is available for free download on *Github* at http://www.goo.gl/gm5vzl [Accessed: 7th May 2014].

5.5 Phonetic feature analysis using *d-prime*

The following sections present a phonetic feature analysis using the signal detection measure *d-prime* (*d'*). This analysis takes into account the results of both consonant identification experiments presented in §5.2 and §5.3, and offers a thorough examination of the types of perceptual errors that participants made. To begin with, the methodology employed for the *d'* analysis is explained, and then the results are discussed separately for the 'quiet' and speech-in-noise data.

5.5.1 Method

The signal detection measure *d'* attributes observers' responses to a combination of response bias (see §5.4.3) and sensitivity (Macmillan & Creelman, 1991). Sensitivity refers to the discriminability of sensory information, and in the present context, to the discriminability of phonetic features encoded in the consonants. The goal is to examine how well the participants detected the presence of consonantal features.[44]

The phonetic features used to specify consonants – and correspondingly the features that participants were required to correctly detect – can be broadly classified into 'manner of articulation', 'place of articulation', and 'voicing' features. As Figure 5.11 shows, the values to characterise the manner of articulation were 'plosive', 'fricative' and 'nasal'. In compliance with the IPA chart (revised to 2005), the place of articulation features were 'bilabial', 'labiodental', 'dental', 'alveolar', 'post-alveolar', and 'velar'. Voicing had two values, namely 'voiced' ('+' = presence of vocal fold vibration) and 'voiceless' ('–' = absence of voicing). As the consonants

[44] The feature analysis neither intends to affirm the psychological reality of features, nor to evaluate different proposed feature sets (see e.g. Jakobson, Fant & Halle, 1963; Chomsky & Halle, 1968; Wang & Bilger, 1973; Keating, 1988; Bernstein *et al.*, 2000; Smits, 2000; Cutler *et al.*, 2004; Clark *et al.*, 2007). The selected features should merely be considered as an analytical instrument used to analyse the consonant confusions on a finer-grained level.

were characterised as either voiced or voiceless, only one category (voicing) will be shown in the subsequent tables and figures.

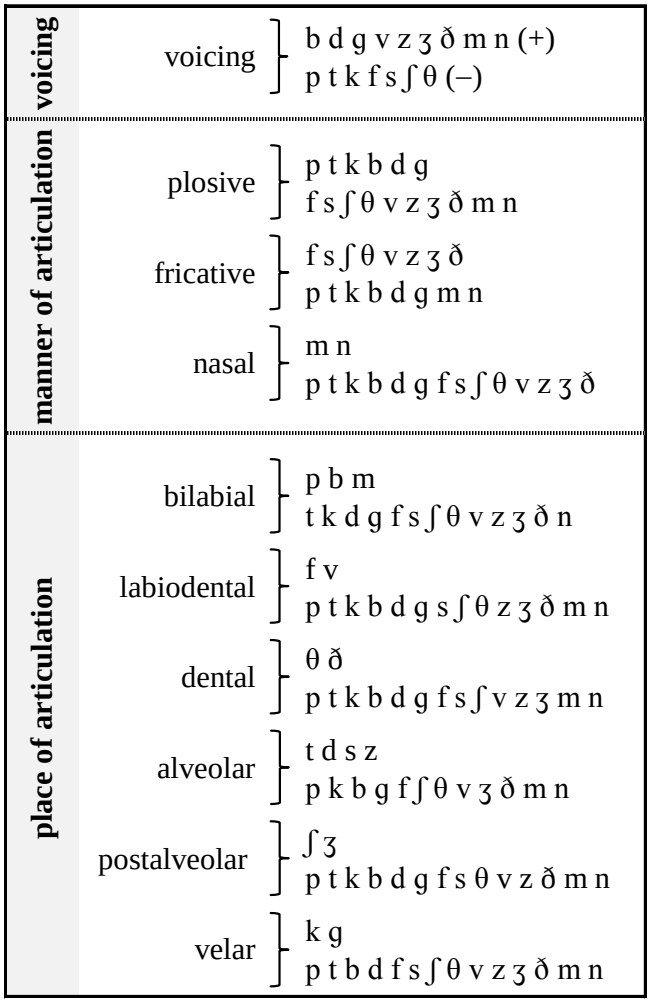

Figure 5.11. Phonetic features used to specify the consonants tested in Experiments 3 and 4. They can be broadly clustered into ‘manner of articulation’, ‘place of articulation’ and ‘voicing’ features. The feature values are shown to the left to the parenthesis, and the corresponding consonants are shown to the right.

A preliminary single-feature and combined-feature analysis (in accordance with Woods *et al.*, 2010) showed that in the quiet listening condition, most errors were single voicing errors (e.g. /ð/ misperceived as /θ/) and single place of articulation errors (e.g. /θ/ misjudged as /f/), followed by combined manner and place of articulation errors (e.g. /m/ misclassified as /v/). This means that the two consonants in a stimulus-response pair only differed with respect to their voicing characteristics,

the place they were articulated at, or both manner and place features. In noise, single place of articulation errors and single voicing errors (e.g. /ð/ misperceived /θ/) occurred most often.

The goal of the *d'* analysis was to describe the patterns observed in the confusion matrices in terms of structural relationships among the consonants across conditions, with a particular view to the impact of facewear on perceptually-relevant information for consonant recognition. The *d'* metric allows us to assess listeners' sensitivity to phonetic features irrespective of a tendency towards the type of perceptual error. This is achieved by taking the covariance of hit (H) and false alarm (FA) rates into account. The FA rate is the proportion of responses for a phonetic feature when a different feature was presented. In other words, FA is the probability that a feature was perceived when it was not actually encoded in the consonant stimulus (incorrect identification). By contrast, H is the probability that a feature was in fact perceived when it was encoded in the stimulus (correct identification). For example, an FA rate of 16% in the control condition (noise/AV) means that in 16% of all trials where the tested consonant was not a plosive, a 'plosive' response was (falsely) given. An H rate of 10% means that in 10% of the cases where the consonant was a plosive, a 'plosive' response was (correctly) given.

From the FA and H scores, *d'* is calculated by subtracting the z-transforms of the FA rates from the z-transforms of the H rates: $d' = z(H) - z(FA)$. The larger the difference between H and FA rates, the higher *d'* will be. A high *d'* value signifies high sensitivity to a particular feature. To illustrate this again with an example, in the rubber mask condition (noise/AV) the place of articulation feature 'labiodental' achieved H = 42% and FA = 7%, whereas 'dental' yielded H = 42% and FA = 21%. Judging from the H rates alone it would seem as if both features were equally well perceived, because they obtained the same proportion of correct responses. However, 'labiodental' has a lower FA rate, for which reason *d'* = 1.3, whereas *d'* for 'dental' would only equal 0.6. Accordingly, the listeners were more sensitive to 'labiodental' than 'dental' in this particular case.

One additional advantage of calculating *d'* over percentage correct or error scores is that sensitivity increases when either H increases or FA decreases (or both).

Furthermore, *d'* is insensitive to the difference in the proportion of consonants that are specified by a certain phonetic feature, compared to the ones that are not. The imbalance in the occurrence of features in the predefined feature set has the effect that a listener with a tendency towards the 'more frequent' case would produce fewer errors than a participant with a tendency to respond with the 'less frequent' case. For example, only two consonants in the current test set were specified as 'dental'. Hence, a consonant was 'less frequently' produced at the dental place of articulation than, say, at the alveolar place of articulation. If a listener was insensitive to 'dental', s/he would perform better if s/he almost never responded with 'dental' ('more frequent' case) than when s/he almost always responded with 'dental' ('less frequent' case). If *d'* didn't take such asymmetries in the feature set into account, the results would reflect more the stimulus material than the observers' perceptions.

5.5.2 Results

Response biases can differ across participants. The *d'* values were for this reason computed in *R* for each participant.[45] After that, the mean *d'* across participants was calculated separately (for the complete table of results see Appendix D.2, Table D.37). The results of these computations are summarised in Tables 5.9 and 5.10.

[45] Thanks to David van Leeuwen for providing the *R* script to run the *d'* analysis.

listening condition	facewear	modality	manner of articulation			place of articulation						voicing
			plosive	fricative	nasal	bilabial	labiodental	dental	alveolar	postalv.	velar	
quiet (Experiment 3)	control	AO	4.3	4.3	5.6	6.2	4.1	3.4	4.3	4.2	4.6	4.4
		AV	4.7	4.7	6.2	5.8	5.2	4.6	4.2	4.0	4.9	3.9
	balaclava (mouth hole)	AO	4.4	4.5	6.0	5.6	4.2	3.7	4.4	4.4	5.4	4.2
		AV	4.3	4.3	6.2	6.2	4.5	3.7	4.2	4.4	5.6	4.0
	tape	AO	4.2	3.6	3.7	3.1	2.9	2.7	2.7	2.9	4.3	3.7
		AV	4.2	3.7	3.8	3.3	3.2	3.1	2.8	2.7	4.1	3.7
	surgical mask	AO	4.4	4.4	5.6	5.4	4.0	3.6	4.4	4.2	4.7	4.0
		AV	4.6	4.6	5.6	5.6	3.9	3.5	4.1	3.9	4.8	4.0
	balaclava (no mouth hole)	AO	4.0	4.1	6.2	5.4	4.1	3.3	3.9	4.1	4.7	4.2
		AV	4.4	4.5	5.6	5.5	4.3	3.6	4.5	4.1	5.5	4.0
	hoodie/scarf	AO	4.7	4.7	6.2	6.2	3.7	3.3	4.4	4.2	5.0	4.0
		AV	4.4	4.4	6.2	5.6	3.9	3.3	4.4	4.3	4.9	4.1
	niqāb	AO	4.6	4.7	6.2	6.0	3.6	3.3	4.1	4.2	5.8	3.8
		AV	4.7	4.8	6.0	6.0	3.7	3.4	4.3	4.4	5.6	3.8
	rubber mask	AO	4.3	4.4	6.2	6.2	3.8	3.4	4.3	4.0	5.0	3.7
		AV	4.2	4.3	6.2	5.1	3.7	3.3	4.2	4.4	5.7	3.7
	helmet	AO	4.7	4.7	6.2	6.2	4.0	3.5	4.4	4.3	5.5	4.2
		AV	4.3	4.2	5.6	5.5	4.4	3.6	4.1	4.2	4.6	3.9

Table 5.9. Results of the *d'* analysis of the perceptual consonant confusion data obtained in Experiment 3, averaged across all participants' individual *d'* results. Darker shading of cells indicates high *d'* values (high detectability of a feature), and lighter shading marks low *d'* values (low detectability). The highest *d'* value of 6.2 denotes perfect recognition (no errors), *d'* = 0 signifies a random response (guessing), and *d'* < 0 suggests a strong response bias (asymmetrical confusion).

listening condition	facewear	modality	manner of articulation			place of articulation						voicing
			plosive	fricative	nasal	bilabial	labiodental	dental	alveolar	postalv.	velar	
speech-in-noise (Experiment 4)	control	AO	1.6	1.8	1.5	1.7	2.1	1.2	1.3	2.0	2.5	1.7
		AV	2.3	3.1	2.2	5.3	4.6	3.9	2.7	2.7	3.7	2.5
	balaclava (mouth hole)	AO	1.2	1.2	1.5	1.5	1.2	0.7	1.3	2.1	1.5	1.2
		AV	1.8	2.3	1.7	4.3	3.5	2.2	1.8	2.8	2.1	1.6
	tape	AO	0.7	0.6	0.3	0.2	0.7	-0.4	0.0	0.6	0.5	0.5
		AV	0.9	0.9	0.2	2.0	2.1	0.8	0.6	0.8	1.1	0.8
	surgical mask	AO	1.3	1.7	1.3	1.4	1.3	1.1	1.1	2.2	1.6	1.5
		AV	1.4	1.7	1.2	1.7	1.6	1.7	1.2	2.1	1.9	1.6
	balaclava (no mouth hole)	AO	1.7	2.0	1.8	1.6	1.7	0.7	1.4	2.5	1.9	1.8
		AV	1.9	2.1	1.8	1.7	1.8	0.9	1.5	2.7	2.2	1.9
	hoodie/scarf	AO	1.7	1.9	1.1	1.5	1.1	1.7	1.4	2.5	1.9	1.4
		AV	1.7	1.9	1.3	1.6	1.0	1.6	1.6	2.5	2.3	1.6
	niqāb	AO	1.4	1.5	0.7	1.5	1.2	0.8	1.0	2.0	2.3	1.1
		AV	1.4	1.6	0.6	1.6	1.1	0.7	1.0	2.1	2.3	1.3
	rubber mask	AO	1.0	0.9	0.9	1.0	1.2	0.8	0.8	1.2	1.1	1.1
		AV	1.0	0.8	0.8	1.0	1.3	0.6	1.0	1.5	1.8	1.0
	helmet	AO	0.3	0.4	0.6	0.2	0.4	0.4	0.0	0.8	0.2	0.4
		AV	0.3	0.5	0.5	0.7	0.3	0.2	0.2	0.7	0.2	0.3

Table 5.10. Results of the *d'* analysis of the perceptual consonant confusion data obtained in Experiment 4, averaged across all participants' individual *d'* results. Darker shading of cells indicates high *d'* values (high detectability of a feature), and lighter shading marks low *d'* values (low detectability). The highest *d'* value of 6.2 denotes perfect recognition (no errors), *d'* = 0 signifies a random response (guessing), and *d'* < 0 suggests a strong response bias (asymmetrical confusion).

During the first series of *d'* computations, cases of perfect recognition of a feature were observed (no errors made by participants). In such cases, *d'* obtained the value of infinity (FA = 0% and H = 100%). The upper H limit was therefore adjusted to 99.9%. Subsequently, a *d'* value of 6.2, which resulted from a hit rate of 99.9% and a false-alarm rate of 0.1%, was considered to represent ceiling performance. As Table 5.9 reveals, this only occurred for 'nasal' and 'bilabial' in the 'quiet' condition.

In Tables 5.9 and 5.10, darker shading of cells indicates high *d'* values, i.e., the corresponding phonetic features were well detected. Lighter shading of cells denotes low *d'* values, i.e., low detectability of the respective features. In cases where FA = H, a *d'* value of zero was assigned (e.g. for 'alveolar' in the helmet/noise/AO condition). $d' = 0$ implies that the participants answered at random, which means that they were insensitive to a particular phonetic feature. $d' < 0$ signifies that a feature caused a strong response bias and was systematically identified as another feature (e.g. 'dental' in the tape/noise/AO condition).

The *d'* results listed in Tables 5.9 and 5.10 are again graphically represented in Figures 5.12 to 5.21. The figures show the *d'* values for each phonetic feature separately. This enables the visual comparison of the detectability of a particular phonetic feature in the two listening conditions (quiet/noise) and modalities (AO/AV), and gives insights into the extent to which the discriminability of the feature changed between the control and facewear conditions. Note that the two balaclavas will appear in the illustrations as 'balaclava 1' (to refer to the balaclava *without* the mouth hole) and 'balaclava 2' (balaclava *with* the mouth hole). This is in keeping with the naming convention used in Chapter 4.

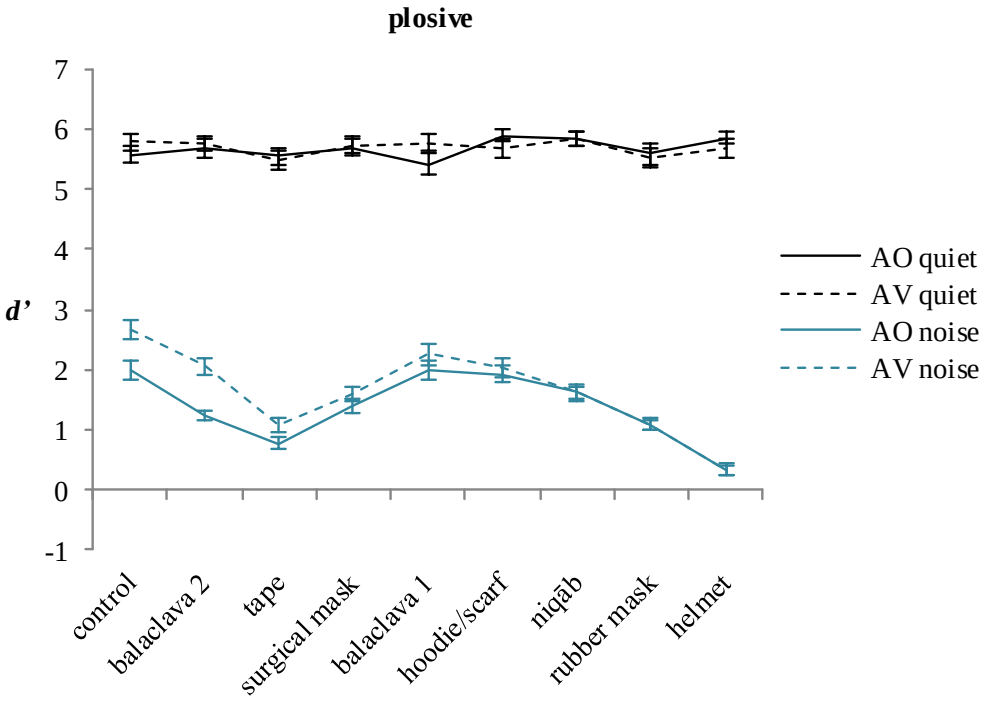

Figure 5.12. Results of *d'* calculations for the manner of articulation feature 'plosive', averaged across participants, for control and each facewear condition separately, as a function of listening condition and modality. The error bars show the standard error of the mean.

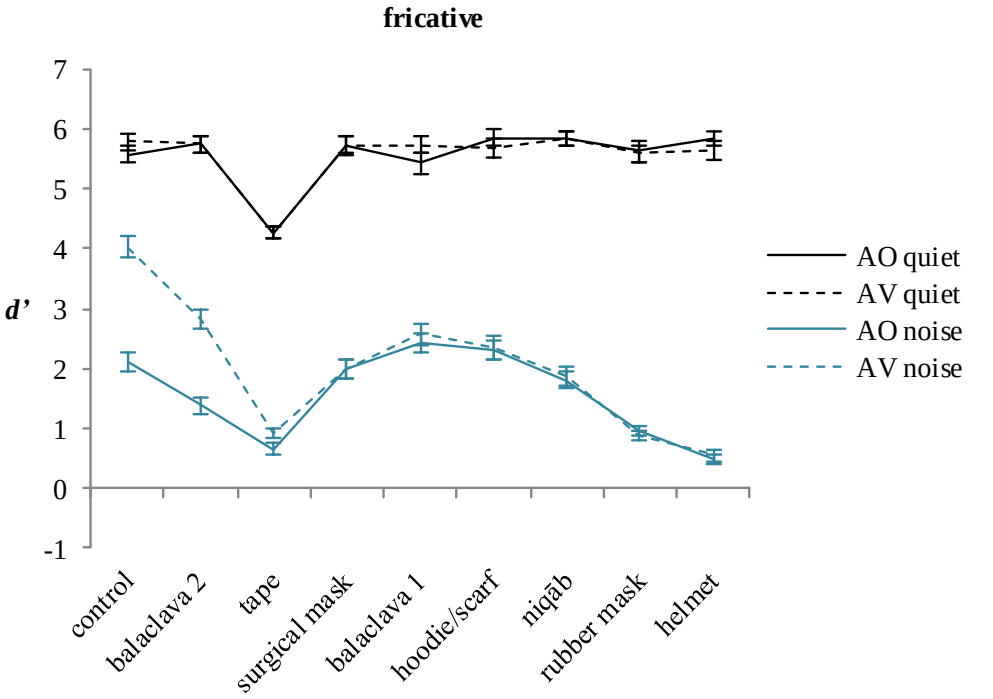

Figure 5.13. Results of *d'* calculations for the manner of articulation feature 'fricative', averaged across participants, for control and each facewear condition separately, as a function of listening condition and modality. The error bars show the standard error of the mean.

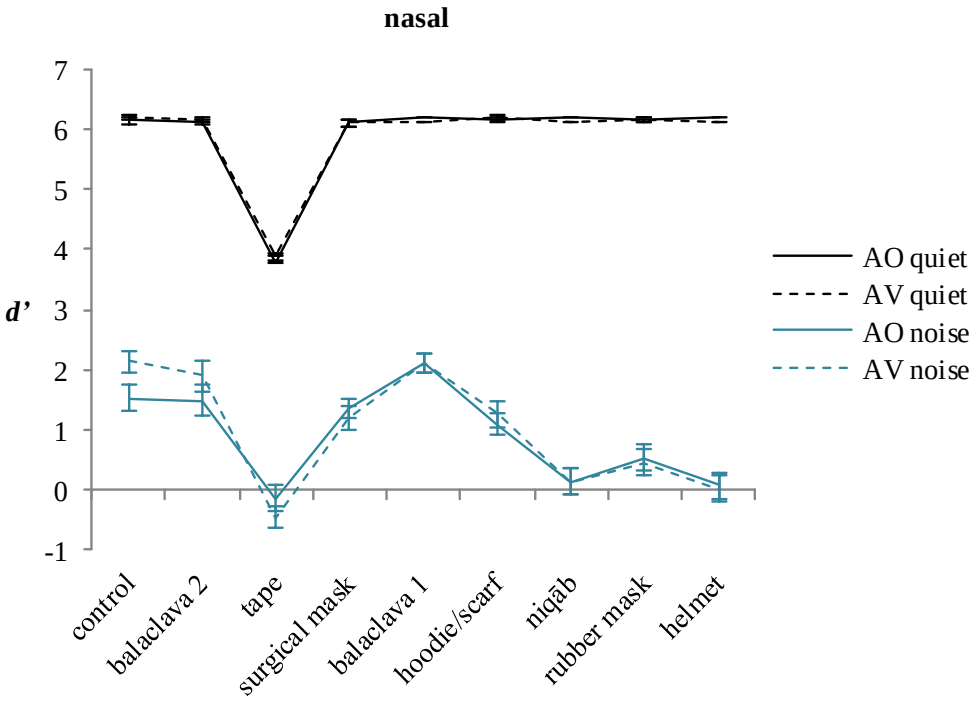

Figure 5.14. Results of *d'* calculations for the manner of articulation feature 'nasal', averaged across participants, for control and each facewear condition separately, as a function of listening condition and modality. The error bars show the standard error of the mean.

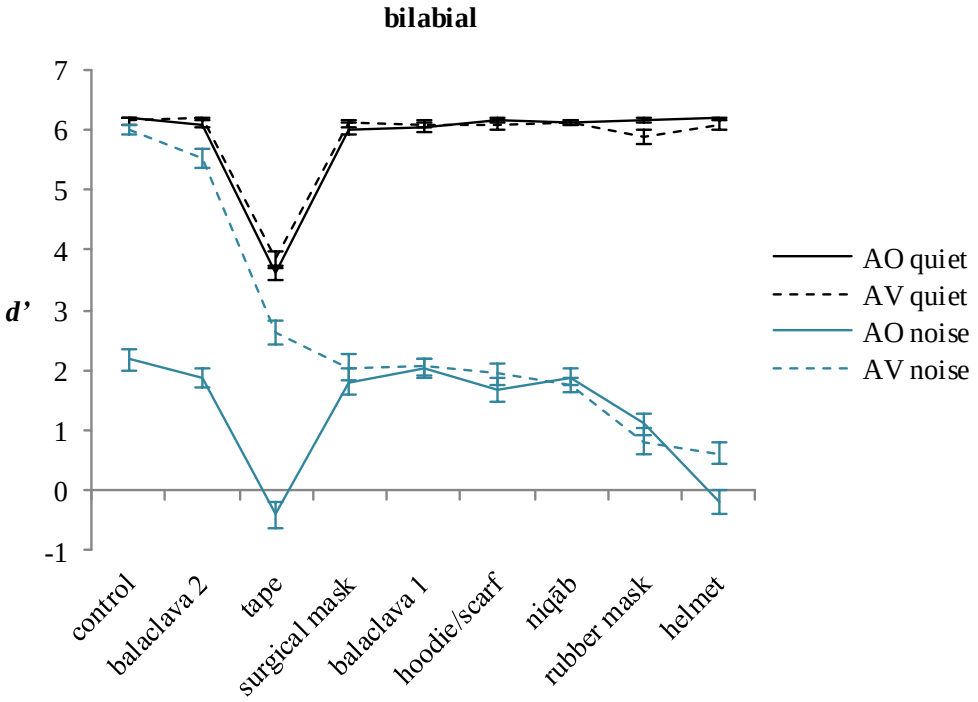

Figure 5.15. Results of *d'* calculations for the place of articulation feature 'bilabial', averaged across participants, for control and each facewear condition separately, as a function of listening condition and modality. The error bars show the standard error of the mean.

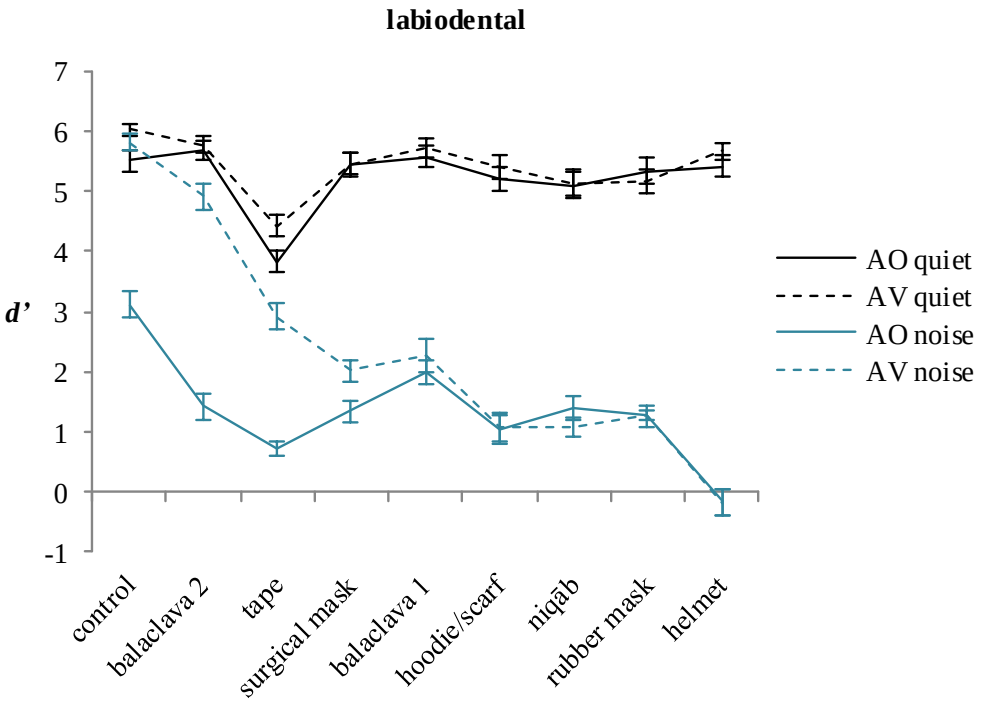

Figure 5.16. Results of *d'* calculations for the place of articulation feature 'labiodental', averaged across participants, for control and each facewear condition separately, as a function of listening condition and modality. The error bars show the standard error of the mean.

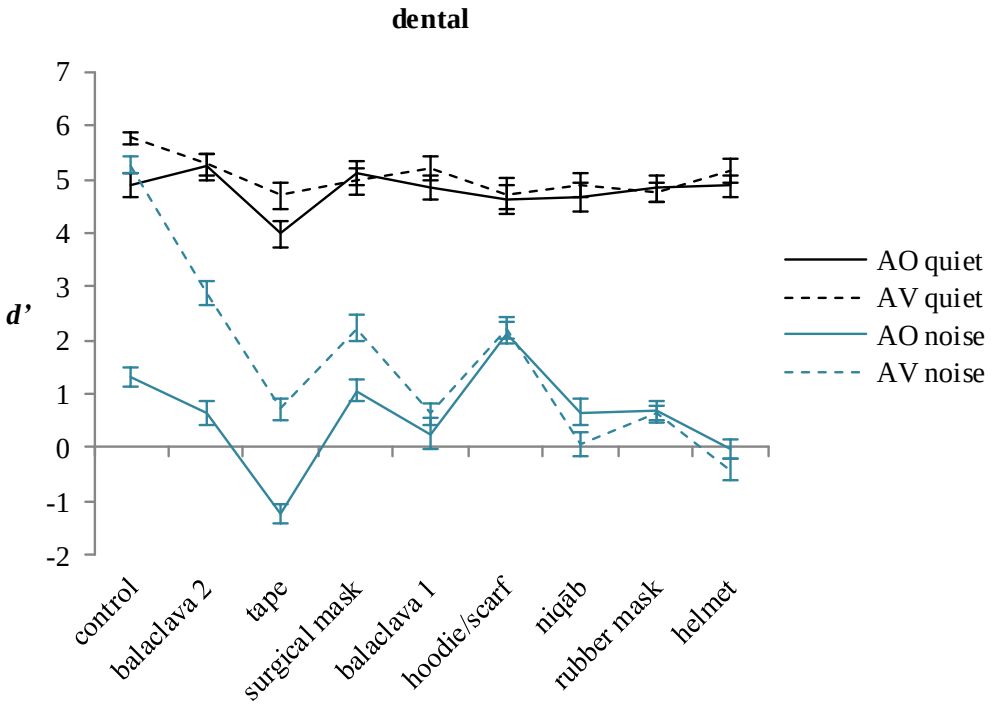

Figure 5.17. Results of *d'* calculations for the place of articulation feature 'dental', averaged across participants, for control and each facewear condition separately, as a function of listening condition and modality. The error bars show the standard error of the mean.

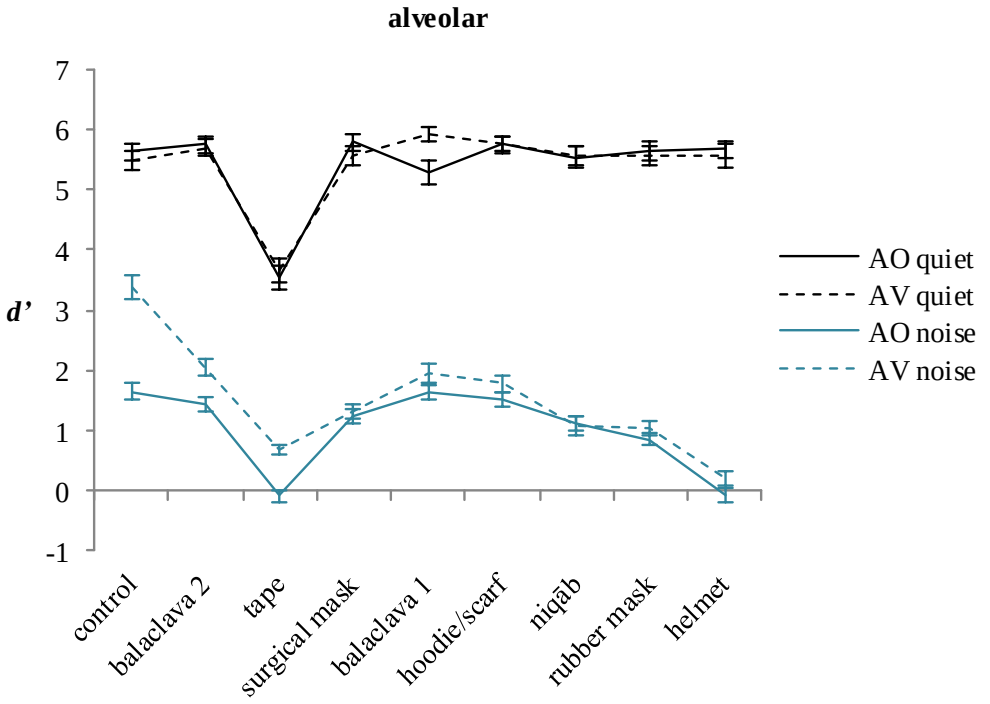

Figure 5.18. Results of *d'* calculations for the place of articulation feature 'alveolar', averaged across participants, for control and each facewear condition separately, as a function of listening condition and modality. The error bars show the standard error of the mean.

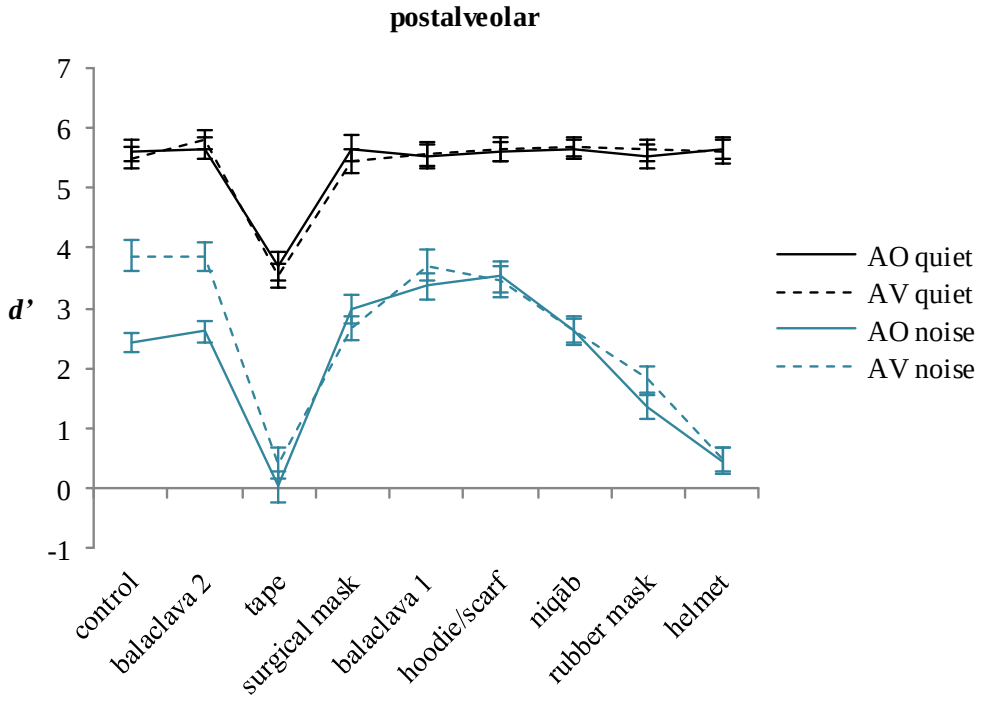

Figure 5.19. Results of *d'* calculations for the place of articulation feature 'postalveolar', averaged across participants, for control and each facewear condition separately, as a function of listening condition and modality. The error bars show the standard error of the mean.

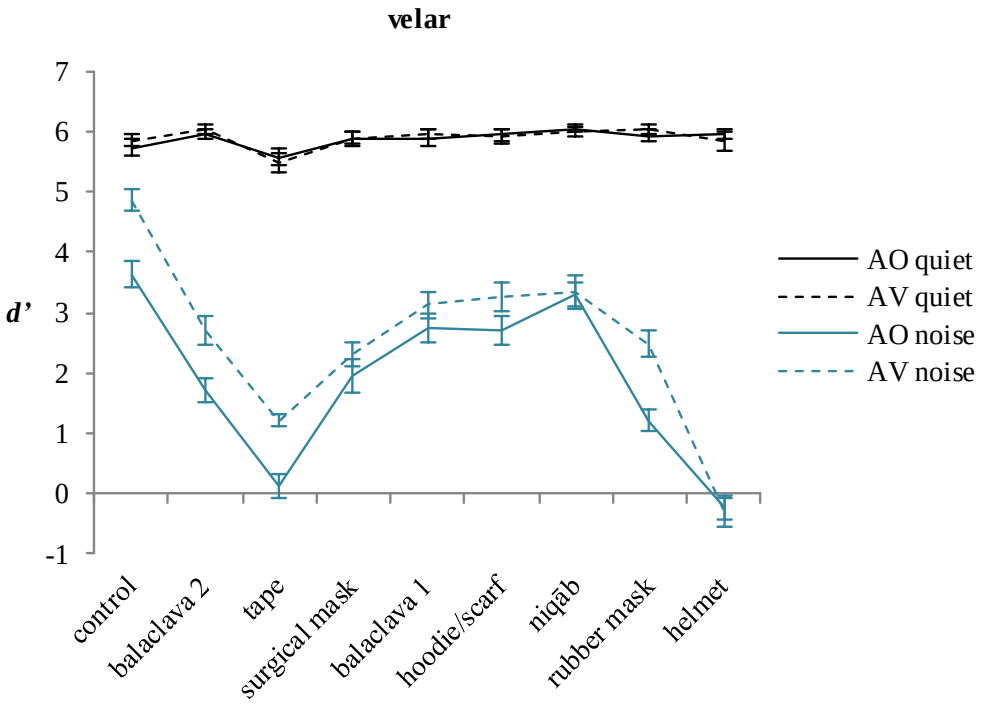

Figure 5.20. Results of *d'* calculations for the place of articulation feature 'velar', averaged across participants, for control and each facewear condition separately, as a function of listening condition and modality. The error bars show the standard error of the mean.

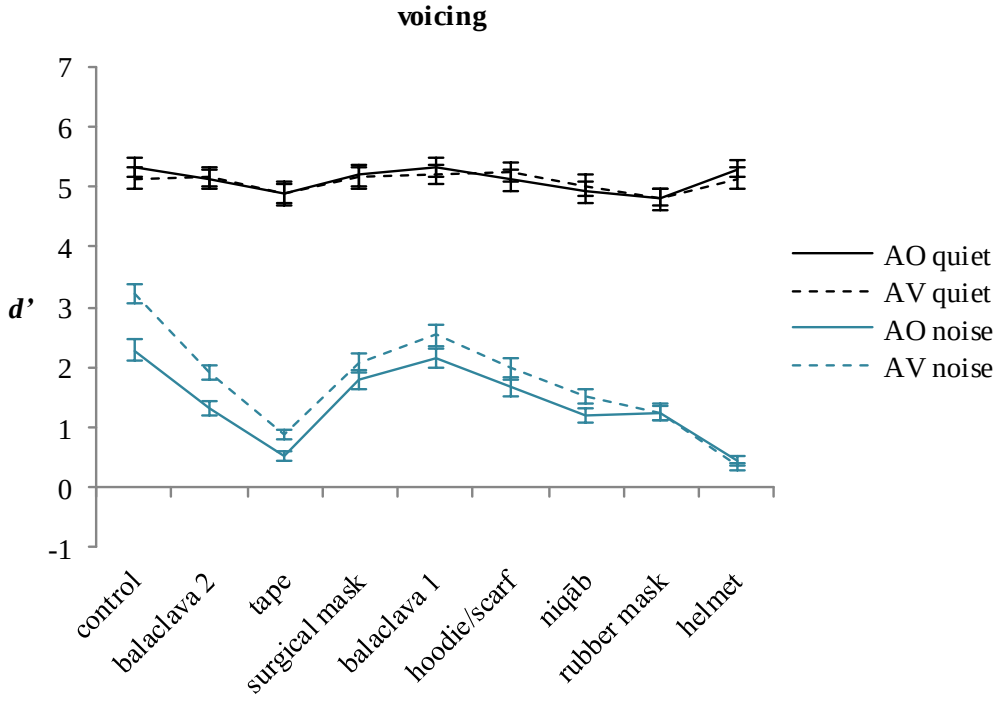

Figure 5.21. Results of *d'* calculations for 'voicing', averaged across participants, for control and each facewear condition separately, as a function of listening condition and modality. The error bars show the standard error of the mean.

Figures 5.12 to 5.21 show that there is a clear discrepancy in sensitivity to phonetic features between the quiet and noise conditions. This recalls the percentage correct scores presented earlier. Sensitivity to features in the quiet listening condition ranged from *d'* = 2.7 (e.g. dental/tape/AO) to *d'* = 6.2 (e.g. nasal/control/AV). Sensitivity in the noise condition ranged from *d'* = –0.4 (dental/tape/AO) to *d'* = 5.3 (bilabial/control/AV). To facilitate comparisons of the identifiabilities of consonant groups which share similar place, manner or voicing features, a series of two-way repeated-measures ANOVAs were carried out. The dependent variables were the *d'* values for the ten phonetic features, and the independent factors were 'modality' and 'facewear'. Results are again reported at $p < .05$. For ease of discussion, the results for the quiet and noise conditions are presented separately.

5.5.2.1 Quiet listening condition (Experiment 3)

In the quiet listening condition, the main effect of modality (AV vs. AO) was significant only for labiodental ($p < .05$) and dental ($p < .01$; see Appendix D.5, Table D.47, for details). This implies that averaged across facewear, only these two features were better detected when visual speech cues were additionally available to participants. The main effect of facewear was significant for all features, which indicates that the participants' ability to detect either of the features was (on average) significantly impaired when the consonants had been produced through facewear. The modality x facewear interaction was significant for bilabial ($p < .01$), dental, and alveolar ($ps < .05$). This suggests a complex interplay between the type of facial disguise and the importance of seeing the talker's face on detecting a certain phonetic feature (for further discussion see §5.6).

To examine the significant interactions further, *post-hoc* Bonferroni-adjusted pairwise comparisons were carried out. These revealed significant differences between AO and AV presentation modality only for some of the features and facewear conditions. In the control samples, only dental ($p < .001$) and labiodental ($p < .01$) were significantly better detected in the AV condition. This means that

observers' sensitivity to (only) these two phonetic features improved when they could see the face during exposure. The same effect was found for dental, labiodental (*p*s < .01), and bilabial ($p < .05$) in the tape condition, for alveolar ($p < .001$) and plosive ($p < .05$) in the balaclava (no mouth hole) condition, and for bilabial ($p < .05$) when the speech was produced through the rubber mask. In all other cases, participants appeared to be equally sensitive to the phonetic information encoded in the stimulus when they only heard the talker's voice, or when they simultaneously heard and saw the talker. Put differently, having supplementary access to visual speech cues did not, in most cases, further improve the detection of phonetic features encoded in the consonants. Participants' performance was for most consonants already near ceiling in the AO condition.

5.5.2.2 Speech-in-noise condition (Experiment 4)

In the speech-in-noise condition, the main effect of modality was significant for all features tested, which again means that (on average) all features were better detected when the participants had access to facial speech cues (for details see Appendix D.5, Table D.48). The main effect of facewear was also significant for all features, indicating that the detection of all consonantal features (averaged across modality) was impaired when the consonants had been spoken through a face covering. The interaction between modality and facewear was significant for plosive, fricative, bilabial, labiodental, dental, alveolar, postalveolar, velar (*p*s < .001), and voicing ($p < .01$). This result once more suggests an extensive interplay between the modality the consonants were perceived in and the type of facewear on the detection of features.

A range of *post-hoc* tests again ascertained the features that were subject to a significant gain in sensitivity in the AV condition (compared to the AO condition). In the control condition, sensitivity to all features was significantly enhanced (*ps* < .001; except nasal, $p < .05$). In the balaclava (mouth hole) condition, sensitivity to all features except nasal increased, i.e., plosive, fricative, bilabial, labiodental, dental, postalveolar, velar (*p*s < .001), alveolar, and voicing (*p*s < .01). Similarly, the

features bilabial, labiodental, dental, alveolar, velar (*p*s < .001), fricative, and voicing (*p*s < .01) were significantly better detected in speech produced through the tape (in the AV condition).

In the surgical mask condition, participants better recognised dental ($p < .001$), labiodental ($p < .01$) and plosive ($p < .05$) in the AV condition than in the AO condition. In addition, sensitivity to plosive and voicing (*p*s < .05) significantly differed between AO and AV for the balaclava (no mouth hole) condition. As for the hoodie/scarf combination, only the place feature velar ($p < .05$) was significantly better detected in the AV condition.

Lastly, in the *niqāb* condition, listeners correctly identified the consonantal features dental and voicing (*p*s < .05) significantly more often in the AV than in the AO condition. In the rubber mask condition, this was only the case for velar ($p < .001$) and bilabial ($p < .05$), and in speech spoken through the helmet, only sensitivity to bilabial ($p < .01$) increased.[46]

[46] Note that Appendix D.3 offers the results of a statistical comparison between the *d'* values obtained in the control condition and the corresponding *d'* values elicited in the facewear conditions.

5.6 General discussion of Experiments 3 and 4

The chapter concludes with a general discussion of the results from both consonant identification perception experiments. The goal of Experiments 3 and 4 was to determine how accurately phonetically-untrained listeners can identify syllable-onset English consonants spoken while the talkers were wearing a variety of forensically-relevant face and head coverings. Participants in the study made consonant judgements during both auditory-only and auditory-visual presentation of the speech stimuli. Owing to the large number of test tokens (576 per participant), a between-group design was adopted. The first participant group was tested with studio quality recordings when the speech stimuli were presented in a quiet listening condition (Experiment 3), and the second group when the original speech was intermixed with 8-talker babble noise at low SNRs (Experiment 4).

Across facewear conditions, a large number of consonant responses (N = 24,768, quiet; N = 22,464, noise) were elicited from a total of 82 participants (N = 43, quiet; N = 39, noise). The primary goal of the study was to estimate how much (if any) visual speech information can still be extracted from the talker's face when crucial articulators are fully or partly disguised. The resultant (predominantly asymmetrical) consonant identification errors were analysed by means of the signal detection measure *d-prime* (*d'*). This aimed to ascertain the extent with which consonantal manner of articulation, place of articulation, and voicing information was transmitted in each facewear condition.

When the speech stimuli (/C_1ɑ:C_2/ syllables embedded phrase-finally in a carrier phrase) were presented in the quiet listening condition, participants (on average) identified 92.2% of the onset consonants (/C_1/) correctly, with hit rates ranging from 94.4% in the most favourable experimental condition (control/AV) to 82% in the least favourable condition (tape/AO). By comparison, consonant recognition accuracy in the speech-in-noise condition was substantially lower throughout. When the speech was embedded in 8-talker babble noise, hit rates markedly declined to 39.2% correct identifications overall, this time ranging from 69% (control/AV) to 12.4% (tape/AO).

A comprehensive analysis of the consonant errors across experimental conditions revealed that fricatives, especially non-sibilants, were particularly difficult to identify. This finding is in line with previous research on human-perceptual consonant recognition, e.g. by Woods *et al.* (2010), Lovitt & Allen (2006), Smits *et al.* (2003), Weber & Smits (2003), Benkí (2003), Redford & Diehl (1999), Wang & Bilger (1973), and Miller & Nicely (1955). Furthermore, most errors made in the quiet listening condition were single voicing and single place of articulation errors, followed by combined manner and place of articulation errors. In noise, single place of articulation and single voicing errors occurred most frequently. The observation that the transmission of consonantal place information is severely disrupted in (auditory) noise corroborates the results reported for auditorily-presented consonants in the above-named studies. Single manner of articulation errors were overall rare in the study, suggesting that consonantal manner is easier to identify than place. This finding is in accordance e.g. with Weber & Smits (2003) and Miller & Nicely (1955). The high number of voicing errors implies that voicing is generally less robust than place information. This accords with Lovitt & Allen (2006), but contrasts with Woods *et al.* (2010), Weber & Smits (2003), Benkí (2003), Wang & Bilger (1973), and Miller & Nicely (1955), who found voicing (along with nasality) to be the most stable consonantal feature in noise.

In the following sections, the results from both experiments are discussed in more detail. The reader's attention is in particular drawn to the finding that the occurrence and strength of the observed auditory-visual effects appear to be related to the type of visual speech information still recoverable from a disguised face, as well as to the specific articulatory and acoustic properties of the tested consonants.

5.6.1 Auditory-visual facewear effects

5.6.1.1 Quiet listening condition (Experiment 3)

When clean speech was presented to the listeners (Experiment 3), a weak but statistically significant gain in consonant intelligibility was observed when visual speech information was presented simultaneously with the soundtrack of the talker's voice (the 'AV effect'). However, when the data were subdivided by type of facewear, a significant AV effect (averaged across consonants) was found only for the tape condition. For all other tested head and face coverings, the participants' recognition accuracy did not significantly differ between modalities. This outcome is less surprising when one bears in mind that the listeners' performance was already very high in the auditory-only condition. The listeners could identify the consonants presented to them vastly above chance level (6%), and in fact performed close to ceiling, even when no video images of the talkers' faces were provided to them. Hence, when the listening conditions were optimal, the presentation of facial speech cues did not further support consonant identification.

The *d'* analysis showed that in fact only a subset of features was better recognised when the talker's (disguised) face was presented in the tape condition. These were the place of articulation features 'bilabial', 'dental', and 'labiodental'. In comparison, despite the absence of an *overall* significant AV effect for speech produced through all other types of face masks, some phonetic features were still better detected when facial cues were present. Specifically, listeners were significantly more sensitive to the place feature 'alveolar' and the manner feature 'plosive' when the talker's face was concealed with a balaclava (no mouth hole), and to the place features 'dental' and 'labiodental' when the face was undisguised (control). The latter finding can be linked to the observed high rate of confusions among dental and labiodental fricatives. The availability of visual speech cues appears to have helped the listeners overcome the difficulties associated with the identification of these types of sounds.

Moreover, the presentation of the talker's face obscured by a rubber mask had a negative effect on the recognition of bilabial sounds (*d'* was significantly lower in the AV than AO modality). Interestingly, the closer inspection of the relevant videos

revealed a McGurk-like effect in some instances. Due to the flexible, rubber-like material and the hole in the mouth region of the mask, the talker's lips could easily be mistaken for the tongue in this case, creating the illusion of tongue tip movement that would be indicative of dental sounds.

5.6.1.2 Speech-in-noise condition (Experiment 4)

Moving on to the speech-in-noise test (Experiment 4), it can firstly be noted that the AO and AV hit rates varied substantially as a function of facewear type, and that significant AV effects were again only found for certain types of masks. The nine facewear conditions (including the control) evenly clustered into three 'classes'. These differed with respect to the occurrence and strength of the AV effect, which in turn could be related to the amount of visual speech information recoverable from the talker's face.

The first class of facewear includes the control condition (absence of facewear), the balaclava with the mouth hole, and the tape across the talker's mouth (see Figure 5.22). The AV effect was strongest in these three conditions. This reflects the findings from earlier studies showing that observers rely more heavily upon speech cues from the face as the listening conditions deteriorate (here, due to background noise). The AV effect was illustrated by the percentage correct scores, which significantly differed between AO and AV presentation modality, and the corresponding *d'* values. The observers' sensitivity to the majority of consonantal features was enhanced when both auditory and visual speech cues were available. The only exceptions were the manner features 'plosive' and 'nasal' and the place feature 'postalveolar' in the tape condition.

The better detection of consonantal features possibly arose from the fact that the lip and tongue movements, as well as many extraoral speech cues (e.g. from the jaw or cheeks), were visible to the participants. Previous studies have shown that under acoustically degraded conditions, visual speech information extractable from the

talker's moving articulators and from facial muscle contractions is relatively stable. As a result, visual speech cues facilitate especially the recognition of the consonantal *place* of articulation (Rosenblum & Saldaña, 1996; Bernstein *et al.*, 2000; Benkí, 2003). The phonetic information available from the talker's mouth region appears to have been of particularly high value to the observers in the present study too. The possibility of lip-reading (extraction of upper/lower lip movements) seems to have greatly aided consonant identification in the control, balaclava (mouth hole) and tape conditions.[47] In the first two of these conditions, tongue motion was additionally visible. The opportunity to extract lip (and in part tongue) movements could explain the high recognition rates in the AV condition, and hence the highly significant AV effects in these three conditions.

control (no facewear) | balaclava (mouth hole) | piece of tape

Figure 5.22. A highly significant 'AV effect' emerged when the talker's face was undisguised (control), concealed with a balaclava (mouth hole), or when the talker's mouth was taped closed. Arguably, this effect was for the most part the result of the talker's mouth region still being visible to the observers, thus enabling lip- and tongue-reading.

By comparison to place information, phonetic information which encodes the *manner* of articulation is not distinct visually (Rosenblum & Saldaña, 1996; Bernstein *et al.*, 2000; Benkí, 2003). That is, the vocal tract changes which contribute to manner distinctions are not visible, or only partly so. For example, it is difficult to

[47] Note that lip-reading was to some extent even possible for the tape, as the product used in this study was a relatively flexible surgical tape which had been slightly loosened from the talker's lips during recording.

detect whether the tongue is completely blocking the air channel to produce the alveolar plosive /t/, or whether it is closely approximating a blockage to generate the homorganic fricative /s/ (Bernstein *et al.*, 2000). The activity responsible for nasality (velum lowering) is completely hidden from view, which is the reason that nasality is visually not informative. This explains why nasality was not better detected when the face was visible (*d'* was equivalent in the AV and AO modality).

Lastly, consonantal *voicing* (vocal fold vibration) was better recognised by the participants when they had access to facial information. This effect was unforeseen (especially in the balaclava condition, where the neck/larynx was occluded). Further experimentation would be necessary to find out more about the extent to which (sub)glottal gestures, which are essentially invisible, correlate with facial movement (see also Burnham *et al.*, 2001).

The second class of facewear includes the surgical mask, the balaclava without the mouth hole, and the hoodie/scarf combination (see Figure 5.23). The statistical analysis again revealed a significant AV effect for these three types of facial disguise, which means that the success in recognising consonants in noise was significantly lower when only the talker's voice was presented. However, the gain in intelligibility in the AV condition was in each case less pronounced than was the case for the facewear in the first class, i.e., the AV effect was overall weaker (larger *p*-values, smaller effect sizes). The *d'* analysis revealed that sensitivity to phonetic features in the AV condition in fact only increased for a small subset of features. These were the place features 'dental' and 'labiodental' and the manner feature 'plosive' in the surgical mask condition, the manner feature 'plosive' and 'voicing' in the balaclava condition, and the place feature 'velar' in the hoodie/scarf condition.

Why was the AV effect weaker when the speech was produced through these three types of facewear? Firstly, all three masks leave the upper face visible, but entirely conceal the talker's mouth, jaw, and neck/larynx (except the surgical mask). Hence, in comparison to the facewear in the first class, the observers could no longer track lip and tongue movements. Secondly, the acoustic filtering effects had a much larger effect on the acoustic speech signal (*N.B.*: This was also the case for the tape.) The fabrics that covered the talker's face are likely to have modified the acoustic signal to

a degree that the changes were perceptually apparent, which consequently reduced consonant intelligibility further.

Figure 5.23. A significant 'AV effect' was observed when the talker's face was disguised with a surgical mask, a balaclava (no mouth hole), or a hoodie/scarf. Here, the entire mouth and jaw region was covered by the mask. However, the facewear was comparatively close fitting, which possibly allowed observers to extract extraoral speech cues and jaw motion.

So why then did consonant identification increase at all when a disguised face was presented to the observers? First of all, it is worth recalling that the (simultaneously presented) acoustic and visual signals were fully aligned (congruent) in this study. Speech intelligibility in noise is generally known to improve when visual speech cues are present because the auditory signal and visible movements of a talker's articulators share temporal, spectral, and spatial properties (Summerfield, 1987; Grant, 2003; Kamachi *et al.*, 2003; Munhall & Vatikiotis-Bateson, 2004). Spatial properties are those which are directly related to the size and location of visual targets (e.g. displacements of the upper and lower inner margins of the lips at midline, or of the area of lip opening; see Grant, 2003). As Grant (2003) points out, when a listener 'watches' a talker speak in a noisy environment, auditory analysis will be directed to the speech signal of interest, which helps to segregate the signal from the surrounding noise. Visual speech cues would inform the listener about when (temporally) to expect peak amplitudes in the acoustic waveform, and where (in the frequency spectrum) to expect these peaks to occur. Hence, the correlated activity between speech kinematics and the acoustic signal assists listeners in extracting the target signal from the noise at SNRs that would otherwise be too low (see e.g.

Rosenblum *et al.*, 1996; Vatikiotis-Bateson *et al.*, 1998; Grant & Seitz, 2000; Kim *et al.*, 2009). The current findings support the notion that the time-varying characteristics of visual speech can be highly informative and play an important role in phonetic perception even for partly or fully disguised faces.

At the outset of this chapter it was reported that listeners perceptually benefit even from rather crude visual speech movements when fine facial detail is absent due to parts of the talker's face being hidden from view. This has been demonstrated most vividly by 'point-light' studies (see §5.1.1.2), which reveal that spatial cues from dynamic point-light displays provide salient information about basic kinematic properties of a talking face, and significantly improve speechreading performance (Rosenblum *et al.*, 1996; Rosenblum & Saldaña, 1996; Jordan *et al.*, 2000). Regarding the facewear in the second class, one such cue which possibly enhanced consonant intelligibility was the 'inflation' of the surgical mask or the scarf caused by the egressive airstream hitting the inner surface of the fabric. This observation resembles the known effect of 'cheek puffing' as an effective visual speech cue (Scheinberg, 1980; Preminger *et al.*, 1998). It could explain, for example, the better detection (higher *d'*) of the manner feature 'plosive' in the AV condition when the consonants were spoken through the surgical mask.

In addition, visual information extracted from the jaw has been shown to be particularly useful to observers. Here, the surgical mask, the balaclava and the scarf wrapped around the talkers' neck/jaw were comparatively close fitting. The talkers' cyclical opening and closing of the jaw could for this reason still be tracked by the participants. The extraction of jaw information may in turn have drawn attention to 'critical events' in the speech signal, such as syllable onsets (Schwartz *et al.*, 2004; Simpson & Cooke, 2005). This is in line with research which has shown that observing jaw gestures helps observers to identify the rhythmic structure, contrastive focus, stress, and emphasis of spoken utterances (Harrington *et al.*, 1995; Dohen *et al.*, 2004a, 2004b; Scarborough *et al.*, 2009).

Such 'visual aids' in identifying syllable onsets are particularly helpful when the target speech is embedded in background noise where *informational* masking is high. This is the case for babble noise, especially where *N* in the *N*-talker babble equals 1

to 8 (Brungart & Simpson, 2005; Simpson & Cooke, 2005; Cooke, 2006; Lecumberri & Cooke, 2006; Cooke *et al.*, 2008; Barker & Shao, 2009). In the present study, 8-talker babble (i.e., a signal composed of speech of 8 talkers) was used to mask the target speech. Work by Simpson & Cooke (2005) has shown that as the number of talkers in the babble increases, so does the number of onsets in the background. Simpson & Cooke suggest that this might divert the listener's attention away from the target speech (i.e., attentional resources are directed at processing the masker rather than the target speech). This will complicate the detection of relevant onsets, and consequently, speech intelligibility will suffer.

Visual cues can help overcome difficulties associated with the adequate allocation of signal energy to the target speech versus the noise masker (Lecumberri & Cooke, 2006). The extraction of jaw movements may have compensated to some extent for the increased number of distracting onsets in the masker used in the present data. Visual speech information extracted from jaw motion may have assisted the listeners to detect relevant onsets (here, the onset of the first consonant in the CVC syllables) even when the face was disguised.[48]

When perceiving speech in noise, the listeners' attention is typically drawn to the mouth and jaw region. Eye-tracking research by Munhall & Vatikiotis-Bateson (1998) has shown that under free viewing conditions participants fixate upon the talker's mouth region significantly more often as the background noise increases. We might speculate that even when the face is occluded (here by means of a face mask), the listener's attention will intuitively be captured by that area.

Finally, the third class of facewear includes the *niqāb*, the rubber mask, and the motorcycle helmet (see Figure 5.24). In these three conditions, the entire face was concealed, except for a small area around the eyes. Even though the recognition of some prosodic cues might still be possible in this case, visual information about the

[48] The 'enhanced syllable onset' criterion was also proposed by Weber & Smits (2003). In contrast to most other related studies, they found that coda consonants embedded in CVC syllables were better recognised than onset consonants. However, syllables were presented in isolation (i.e., without a carrier phrase) in their study. The authors argue that the moment of stimulus onset was therefore much more uncertain, for which reason the listeners' performance for the onset consonants was reduced.

segmental content of speech is no longer available to observers (or at least massively compromised). It seems that the facewear in this class allowed neither lip-/tongue-reading nor the extraction of any other relevant facial movements from the face. It is for this reason perhaps unsurprising that no AV effects were found. A (more or less) fully-concealed face will, naturally, provide no facial information which would enhance speech perception on the segmental level.

niqāb (full-face veil) | rubber mask | motorcycle helmet

Figure 5.24. No 'AV effect' was registered when the talker's face was concealed with a *niqāb*, a rubber mask, or a motorcycle helmet. Here, no or only very few visual speech cues could be extracted from the talker's articulating face, for which reason consonant intelligibility was not enhanced when the face was presented.

5.6.2 Summary

In conclusion, then, the present study established consonant identification accuracy scores for 'quiet' speech and speech embedded in noise. These were obtained from phonetically-untrained observers who participated in an auditory-only (AO) and auditory-visual (AV) consonant identification experiment where the talker's face had been obscured by one of eight types of face coverings.

The study extends previous research on AO and AV speech perception in quiet and noisy conditions, and offers new insights into the effects of realistic facial occlusions on consonant identification. In contrast with preceding research, which mainly examined the relevance of carefully-defined facial areas during AV speech

processing, this study enhanced the naturalness of the AV speech material by testing a fairly large variety of face/head coverings which are routinely, and in comparatively uncontrolled ways, encountered in real-life communicative situations.

The main findings can be summarised as follows:

- perceptual properties of syllable-onset consonants are changed when these are produced while the talker's face is concealed by facewear
- phonetically-untrained observers are better at identifying consonants when they can also see the talker's articulating face, as opposed to when they only listen to the talker's voice ('AV effect')
- the magnitude of the changes to speech perception, and the type of facial speech cues which support consonant intelligibility, vary greatly with facewear type
- quiet listening condition
    - highly accurate consonant identifications despite facewear (92.3% correct)
    - overall weak but statistically significant gain in consonant intelligibility when visual speech cues are available to observers (see Table 5.11)
    - statistically significant drop in AO and AV consonant intelligibility (compared to the control condition) only in the tape condition
- speech-in-noise condition
    - lower mean consonant identification performance (39.2% correct)
    - considerable AV effect across facewear conditions, signifying that observers start to rely much more heavily upon visual speech cues from the talker's face as listening conditions deteriorate (see Table 5.11)
    - significant drop in AV consonant intelligibility (compared to the baseline) in all facewear conditions, and in AO consonant intelligibility in the tape, rubber mask, helmet, *niqāb*, and balaclava (mouth hole) conditions
- visual speech cues can be recovered even from a partly or fully disguised face
    - strongest AV effect when lip- and/or tongue-reading possible (*cf.* control, balaclava with mouth hole, strip of adhesive tape across the mouth)

- weaker AV effect when mouth region obscured by facewear (*cf.* surgical mask, balaclava without mouth hole, hoodie/scarf combination)
- no AV effect in absence of visual speech cues (*cf. niqāb*, rubber mask, motorcycle helmet)
- perceivers make effective use of extraoral facial cues to consonant identity (e.g. mask 'inflations' or cyclical opening and closing of the jaw, which support syllable onset recognition)

% correct consonant identification				
	quiet listening condition		**speech-in-noise**	
facewear	**AO**	**AV**	**AO**	**AV**
control (no facewear)	93.5	94.4	47.8	69.0†††
balaclava (mouth hole)	93.7	94.2	40.4*	56.8†††***
tape	82.0***	84.2†***	12.4***	27.2†††***
surgical mask	93.7	93.1	43.3	47.3††***
balaclava (no mouth hole)	93.0	94.2	45.0	48.3††***
hoodie/scarf	93.1	93.6	46.6	49.8†***
niqāb	92.4	92.7	40.1**	41.0***
rubber mask	92.5	92.3	32.0***	32.8***
helmet	94.0	93.6	12.8***	12.5***
mean	**92.0**	**92.5**†	**35.6**	**42.7**†††

Table 5.11. Consonant identification accuracy averaged across consonants, for each listening condition (quiet = Experiment 3, noise = Experiment 4) and facewear condition (including control) separately, as a function of modality. '†††' denotes a significant 'AV effect' at $p < .001$, '††' at $p < .01$, and '†' at $p < .05$. '***' denotes a significant difference from the corresponding control condition at $p < .001$, '**' at $p < .01$, and '*' at $p < .05$.

With the results from the two consonant identification experiments in mind, the following chapter again looks at the perceptual characteristics of consonants produced through facewear. However, the study presented in Chapter 6 goes one step further and examines the perceptual properties of consonants which provide an indication of the talker's 'identity'. Specifically, listeners are tested for their ability to correctly determine whether two short consonant-vowel utterances originate from the same talker, or whether they were produced by two different individuals.

6 Talker discrimination based on facewear speech

6.1 Introduction

Chapter 5 dealt with the ability of lay listeners to auditorily and auditory-visually identify a set of consonants when these had been produced while the talkers were wearing facewear. The focus of the study was hence on the *content* of the speech. By contrast, the study discussed in the present chapter calls attention to the *indexical* (talker-specific) properties of speech. Here, it is investigated whether lay listeners can successfully distinguish between two unfamiliar talkers, i.e., whether they can determine if two short samples of speech (/Cɑː/ syllables) with systematically-varying consonantal content (/t p s f n m/) were spoken by the same talker or by two different talkers. The study explores whether a) the listeners' performance in the task is reduced when their decisions are based on facewear speech, and b) some consonants bring about higher talker discrimination rates than others. The latter aspect is based on previous research, which is introduced in the next section. The research questions are:

- Can lay listeners correctly determine whether two samples of speech originate from the same talker or from two different talkers when all the listeners have available for comparison are short CV syllables?
- Does facewear change the talker-specific properties of speech? Specifically, is there any impairment to talker discrimination based on individual consonants and vowels when the speech sounds have been produced while the talker's face is disguised by facewear? In other words, does facewear negatively impact on talker discriminability?
- Does the segmental content of the tested speech samples (here, six different consonants) have an effect on the listeners' performance in distinguishing between unfamiliar talkers?[49]

[49] Some of the results of this study were presented in 2014 at the *23rd Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*.

6.1.1 Speech content and indexical information

According to conventional accounts, there exist two separate mechanisms for the processing of the content of an utterance – hereafter referred to as 'segment processing' – and for the processing of talker-specific (indexical) information extracted from a talker's voice and speech patterns. The latter, for ease of discussion, is henceforth termed 'voice processing'.[50]

The recognition of talkers based on indexical information encoded in the speech signal has often been considered as quite separate from extracting the linguistic content of an utterance (Nygaard *et al.*, 1994). Many of the early theoretical accounts of speech perception propose that during segment processing, the speech input undergoes a normalisation process by which the listener extracts and discards talker-specific properties of the signal (Ladefoged & Broadbent, 1957; Abercrombie, 1967; Ladefoged, 1967; Laver & Trudgill, 1979; Liberman & Mattingly, 1985; Fowler, 1986; McClelland & Elman, 1986; Johnson, 1990; Nygaard *et al.*, 1994; Pisoni, 1997; Remez *et al.*, 1997; Yakel *et al.*, 2000; Rosenblum, 2005).

However, a wealth of evidence from recent behavioural and neurological research suggests that indexical and segmental information are not independent in perception, but interact at an early stage of processing (Bricker & Pruzansky, 1966; Johnson, 1990; Mullennix & Pisoni, 1990; Knösche *et al.*, 2002; Andics *et al.*, 2007; Kraljic & Samuel, 2007). This interdependence of voice and segment processing is illustrated in Figure 6.1. The figure aims to highlight the finding that a processing dependence can occur in both directions (indicated by the arrows). This means that phonetic information about the speech content can influence voice processing, and phonetic

[50] The author acknowledges that the comparison of speech recordings of two individuals does not only involve the analysis of the talkers' voices, but also aspects of their speech which concern the language and/or non-linguistic behaviour (see e.g. French *et al.*, 2010). In the present context, the term 'voice processing' was chosen for the sake of convenience, and in keeping with much of the psychology, psycholinguistic and cognitive literature. Here, a 'voice' is often rather broadly attributed to the auditory percept of vocalisations of a human individual which can be used to recognise the individual. This includes all linguistic and non-linguistic aspects of the vocal signal produced by the talker (see Andics, 2013: 10ff.).

information about talker-specific details encoded in the signal can affect segment processing. These notions will be explained further in the following sections.

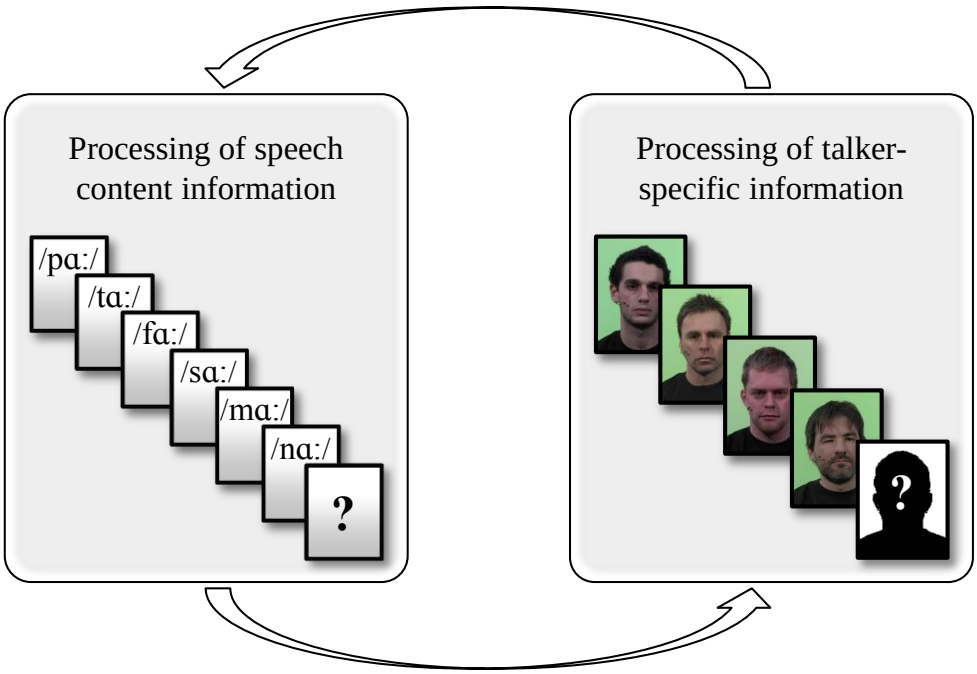

Figure 6.1. Interdependence of processing of speech content and processing of indexical (talker-specific) information encoded in the speech signal. Research has shown that phonetic information about the content of a linguistic utterance (here, segmental content) can influence 'voice processing', and phonetic information about talker-specific details can affect 'segment processing'.

On the one hand, studies have shown that the indexical properties of the signal are *not* entirely discarded during segment processing. Rather, the success of determining the linguistic content of an utterance appears to be partly dependent on talker-specific information. Nygaard *et al.* (1994) note that both types of information encoded in the speech signal become part of a rich and highly detailed representation of the talker's utterance. The perceptual and memory processes involved during speech recognition are hence likely to be affected one way or another when the listener is acquainted with talker-specific phonetic/linguistic detail (through perceptual learning).

Indeed, a large number of researchers have demonstrated that the recognition of speech content is facilitated and that recognition memory is enhanced when the

listener has experience (i.e., is familiar) with the talker's voice and speech patterns. For example, words produced by familiar talkers (or by talkers with voices that are perceptually similar to that of a familiar talker) are generally better identified than words produced by non-familiar talkers. Moreover, it has been found that even individual phonemes are better recognised when presented in single- as opposed to multiple-talker lists (Mullennix & Pisoni, 1990; Nygaard *et al.*, 1994; Goldinger, 1996; Pisoni, 1997; Goldinger, 1998; Nygaard & Pisoni, 1998; Yakel *et al.*, 2000; Lachs & Pisoni, 2004; Sheffert & Olson, 2004; Goh, 2005; Rosenblum *et al.*, 2007; Winters *et al.*, 2008; Davis & Kim, 2010; Cutler *et al.*, 2011).

On the other hand, research has shown that voice processing can significantly draw on the content of an utterance. Hence, the successful extraction of talker-specific information from the speech signal seems to be in part dependent on the speech content (Bricker & Pruzansky, 1966; Mullenix & Pisoni, 1990; Fellowes *et al.*, 1996; Remez *et al.*, 1997; Schiller *et al.*, 1997; Andics *et al.*, 2007; Winters *et al.*, 2008; Cutler *et al.*, 2011). In one of the earliest accounts of this effect, Bricker & Pruzansky (1966) report that listeners' success in identifying talkers varied with the content of the presented speech samples (especially with vowel type). Furthermore, the frequently-cited study by Remez *et al.* (1997) provided evidence that familiar talkers can be correctly identified even when acoustic attributes of voice quality and other non-segmental indexical information have been eliminated. The authors presented (intelligible) sinewave replicas of natural speech to lay listeners, and found that talker identification was at a comparable level to talker identification based on natural speech. They inferred from this result that listeners seem to have access to sufficient talker-specific information for making accurate decisions about talker 'identity', as long as the idiosyncratic segmental variation of speech is preserved.

More recently, Andics *et al.* (2007) studied the effects of segmental information on talker discrimination.[51] Talker discrimination involves the comparison of two samples of speech which were produced either by the same talker or by two different

[51] The expression 'talker discrimination' is used synonymously with 'voice discrimination' in the relevant literature (e.g. Bricker & Pruzansky, 1966; Kreiman & Papcun, 1991; Nygaard *et al.*, 1994; Nygaard & Pisoni, 1998; Winters *et al.*, 2008; Davis & Kim, 2010). In the context of this thesis, the term 'talker discrimination' is given preference.

individuals, and making the decision about whether the samples originated from the same talker or not. The experiment by Andics *et al.* (2007) tested how accurately a group of untrained listeners could distinguish between 13 male talkers based on isolated CVC words with systematically-changing segmental content (e.g. [mɛt], [mɛs], [lɛs], [lɛt]). In each trial of the experiment, participants were asked to decide whether a word had been spoken by the same or by a different talker as the preceding word ('same/different forced-choice one-back' procedure). The authors then compared the responses as a function of the segmental content of each word, and found that some segments led to better talker discrimination than others. Overall, 87.2% of talker discriminations were accurate. There was a higher rate of correct responses for words which contained an onset /m/ as opposed to an onset /l/, a nucleus /ɛ/ versus a nucleus /ɔ/, and a coda /s/ compared to a coda /t/.

In a follow-up study, Cutler *et al.* (2011) confirmed once more that the processing of voice and segment information is interdependent, i.e., that processing dependence emerges in both directions. They tested listeners' performance in talker discrimination based on VC syllables. Here, participants were familiarised with the voices of two male talkers, and in each experimental trial had to categorise the talkers as either 'Peter' or 'Thomas' ('two-alternative forced-choice' procedure). They did so while the syllable content was either constant (always [ot]) or varied ([ɛt], [ɛs], [ot], and [os]). This task aimed to test the effect of segment variation on voice processing.

In addition, the researchers tested the influence that a change in talker has on segment classification (consonant and vowel decisions) when the talkers producing the speech sounds were either constant (always Peter) or varied (Peter or Thomas) within each experimental trial. This task was designed to explore the effect of voice/talker variation on segment processing.

Cutler *et al.* observed a significant 'Garner effect' (Garner, 1974). In this context, this means that the participants' responses were significantly slower when the non-target dimensions – i.e., 'segment type' in the voice processing task, and 'voice type' in the segment processing task – varied compared to when they were constant within a trial. The higher error rates and response times in the talker discrimination

compared to the segment classification task suggest that the impact of segment variation on talker 'identity' decisions was even stronger than the influence of talker variation on speech content decisions.

6.1.2 Aim of the study

The current study builds on the findings by Andics *et al.* (2007) and Cutler *et al.* (2011), who report that some consonants and vowels help lay listeners to discriminate between talkers more than others. The experiment once again centres on the question of whether two types of phonetic information encoded in the speech signal – that is, information about the speech content and information about indexical properties of the speech – are processed independently or in a way that would suggest that they are dependent on one another.

The focus of the study will be on the perception of six consonants embedded in /Cɑː/ syllables which were elicited from four male talkers, all of whom were unfamiliar to the listeners. On the basis of the findings from the aforementioned research, it is hypothesised that the listeners' ability to correctly distinguish between two talkers will vary across the six consonants. It is anticipated that some consonants will carry a greater amount of talker-specific information than other consonants in the test set, and will hence lead to higher correct talker discrimination rates than others.

In keeping with the scope of the thesis, Experiment 5 additionally examines the extent to which facewear affects the listeners' performance in distinguishing between the speech of two unknown individuals. The question that arises is whether the ability of lay listeners to successfully discriminate between talkers based on short speech samples will be further complicated when the speech material was produced while the talker's face/mouth was occluded by facewear. Put another way, does facewear impact on talker discriminability?

The two types of facewear included in the experiment are the motorcycle helmet and the piece of tape adhered to the talker's mouth/cheeks (the reasons for this selection

are given in §6.2.1.2). Based on the findings from the empirical studies presented in Chapters 4 and 5, namely that facewear has the potential to considerably alter certain acoustic and auditory-perceptual properties of consonants, it is hypothesised that facewear will negatively affect unfamiliar talker discrimination. That is, it is expected that talker discrimination based on facewear speech will be more difficult for the listeners than talker discrimination based on control speech.

6.2 Experiment 5: Talker discrimination

The upcoming sections report on the methodology applied to address the research questions raised in the introduction to this chapter. Following this, the results of a statistical analysis of the perception data obtained in Experiment 5 are presented.

6.2.1 Method

6.2.1.1 Participants

Twenty-four participants (13 females, 11 males) were recruited at the MARCS Institute, University of Western Sydney, Australia.[52] Their mean age was 25.2 years (*SD* = 5.1), and none of them reported a history of hearing impairment. The majority were native Australian English speakers, with very few having a bilingual background. All participants had prior knowledge of the study of psychology, and some of them had an understanding of linguistics, phonetics, and psycholinguistics, but none of them had had extensive formal ear training or experience with phonetic analysis. Moreover, no participant reported previous experience of wearing any type of facewear, or interacting with people who do so, on a regular basis. All volunteers participated in the 1-hour experiment in return for a small remuneration.

[52] This work was conducted in 2012 during the author's secondment at the MARCS Institute, University of Western Sydney, Australia, as part of her contractual obligation as a member of the Marie Curie Initial Training Network 'Bayesian Biometrics for Forensics (BBfor2)'.

6.2.1.2 Speech material

The speech material was again extracted from the AVFC corpus (see Chapter 3). The data were taken from four male talkers, who were judged as having the most similar-sounding voices. The average age of the talkers was 28.8 years (*SD* = 7.4). All 24 participants in the perception experiment were unfamiliar with the four talkers prior to taking part in the study.

The test material was extracted from the CVC nonsense syllables and consisted of CV syllables only. This was intended as a way to limit the speech available to the listeners to an even greater extent, and to ensure the same phonetic content per experimental trial (details given below). Specifically, the /C_1ɑːC_2/ nonsense syllables were truncated to open syllables by manually excising the coda consonant using *Praat 5.3.24*. To recall, the /C_1ɑːC_2/ syllables with the same consonantal onset (/C_1/) each had a different coda (/C_2/). The syllables that had a nasal in coda position were excluded here so as to avoid marked anticipatory coarticulation effects. The /Cɑː/ syllables that were tested in this study included six consonants, namely the voiceless fricatives /f/ and /s/, the voiceless plosives /p/ and /t/, and the (voiced) nasals /m/ and /n/. These were all consistently followed by the open back vowel /ɑː/, and were presented without the carrier phrase in which they had originally been uttered. The choice of fricatives and plosives was motivated by the acoustic experiments discussed in Chapter 4. The choice of nasals was based on previous studies which had shown that nasals can carry a high amount of talker-specific information (Nolan, 1997; Amino & Arai, 2009; Kavanagh, 2013).

Of the eight types of facewear included in the AVFC corpus, only two were chosen for the experiment (this being necessary to constrain the length of the experiment). These were the motorcycle helmet and the tape across the talker's mouth. This selection was based on, firstly, the author's experience with forensic phonetic casework in which these two forms of facewear were of concern (see §1.1.2.1). Secondly, the experiments presented in previous chapters have demonstrated that the adverse effects on selected acoustic properties of the speech signal, and also the detrimental perceptual effects that relate to them, were by and large most pronounced for these two types of facewear. To provide a baseline against which the results from

the facewear conditions could be compared, the study also included the control condition (no facewear).

Finally, in line with the empirical work presented so far, the digital audio recordings used here were the ones made with the DPA headband microphone in its original format (48kHz, unfiltered).

6.2.1.3 Stimulus design

The study tested three facewear conditions, namely ‘control’, ‘helmet’ and ‘tape’. The degree of talker discriminability across conditions was measured by means of a ‘two-interval forced-choice (2IFC)’ procedure (see e.g. Kim & Davis, 2003; Davis & Kim, 2006). In each trial of the experiment two pairs were presented serially, i.e., ‘pair 1’ followed by ‘pair 2’ (see Table 6.1). Each pair consisted of two samples (‘sample 1’ and ‘sample 2’) of /Cɑː/ syllables produced either by the same talker (e.g. AA’) or by two different talkers (e.g. AB’). Sample 1 of each pair was always the same token spoken by the same talker, and was hence the standard against which sample 2 in the pair could be judged. However, participants in the experiment were not informed of this characteristic of the stimuli.

The consonants were kept constant across trials, which meant that participants consecutively listened to the same type of /Cɑː/ syllable four times (e.g. /tɑː/–/tɑː/ + /tɑː/–/tɑː/). In the helmet and tape conditions, sample 1 of each pair always consisted of the token that was recorded while the talker’s face/mouth was occluded by the motorcycle helmet or the tape (represented in bold letters in Table 6.1), whereas sample 2 consisted of the token recorded without facewear. In the control condition, both sample 1 and sample 2 of each pair were tokens produced without facewear.

control		helmet		tape	
order 1					
pair 1 (same talker)	**pair 2** (different talker)	**pair 1** (same talker)	**pair 2** (different talker)	**pair 1** (same talker)	**pair 2** (different talker)
A A’	A B’	**A** A’	**A** B’	**A** A’	**A** B’
A A’	A C’	**A** A’	**A** C’	**A** A’	**A** C’
A A’	A D’	**A** A’	**A** D’	**A** A’	**A** D’
B B’	B A’	**B** B’	**B** A’	**B** B’	**B** A’
B B’	B C’	**B** B’	**B** C’	**B** B’	**B** C’
B B’	B D’	**B** B’	**B** D’	**B** B’	**B** D’
C C’	C A’	**C** C’	**C** A’	**C** C’	**C** A’
C C’	C B’	**C** C’	**C** B’	**C** C’	**C** B’
C C’	C D’	**C** C’	**C** D’	**C** C’	**C** D’
D D’	D A’	**D** D’	**D** A’	**D** D’	**D** A’
D D’	D B’	**D** D’	**D** B’	**D** D’	**D** B’
D D’	D C’	**D** D’	**D** C’	**D** D’	**D** C’
order 2					
pair 1 (different talker)	**pair 2** (same talker)	**pair 1** (different talker)	**pair 2** (same talker)	**pair 1** (different talker)	**pair 2** (same talker)
A B’	A A’	**A** B’	**A** A’	**A** B’	**A** A’
A C’	A A’	**A** C’	**A** A’	**A** C’	**A** A’
A D’	A A’	**A** D’	**A** A’	**A** D’	**A** A’
B A’	B B’	**B** A’	**B** B’	**B** A’	**B** B’
B C’	B B’	**B** C’	**B** B’	**B** C’	**B** B’
B D’	B B’	**B** D’	**B** B’	**B** D’	**B** B’
C A’	C C’	**C** A’	**C** C’	**C** A’	**C** C’
C B’	C C’	**C** B’	**C** C’	**C** B’	**C** C’
C D’	C C’	**C** D’	**C** C’	**C** D’	**C** C’
D A’	D D’	**D** A’	**D** D’	**D** A’	**D** D’
D B’	D D’	**D** B’	**D** D’	**D** B’	**D** D’
D C’	D D’	**D** C’	**D** D’	**D** C’	**D** D’

Table 6.1. The stimulus design in which the letters A, B, C, and D each represent speech tokens spoken by four different talkers. There were three facewear conditions (control, helmet, tape). In each trial two pairs of speech samples were presented (pair 1, pair 2). Participants were required to judge which pair consisted of speech produced by the same talker. In the helmet and tape conditions, sample 1 in each pair always consisted of the token produced through facewear (represented by bold/coloured letters), whereas sample 2 consisted of the token recorded without facewear. Two sets of stimuli were prepared across which the order of the same- and different-pairs was counterbalanced (order 1, order 2).

Two sets of stimuli were prepared across which the order of pair 1 and pair 2 in a trial was counterbalanced. In the first set, pair 1 consisted of the speech tokens of the same talker and pair 2 contained the speech tokens of different talkers (‘order 1’). In the second set, this order was reversed, such that pair 1 consisted of the different-talker tokens and pair 2 contained the same-talker tokens (‘order 2’). The aim of this was to control for a potential response bias on the part of the listeners, i.e., to compensate for the possibility that the listeners would favour the first or second pair response across experimental trials.

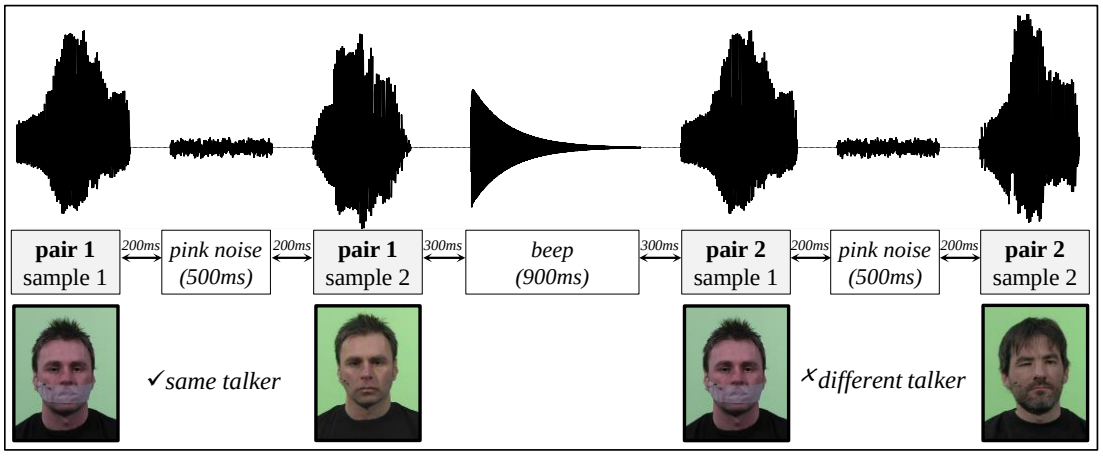

Figure 6.2. Schematic representation of one experimental trial of Experiment 5. The durations of the pink noise and the beep, as well as the interstimulus intervals, were kept constant across trials. All sound files were normalised for amplitude (samples at 70dB, pink noise at 50dB, beep at 60dB). The samples in the same-talker pairs were extracted from different /C_1ɑ:C_2/ syllables so as to avoid the possibility that the listeners’ responses were based on auditory change detection rather than speech processing. The first sample in each pair was always spoken by the same talker in the same facewear condition. Note that the images merely aim to illustrate the experimental design; they were not shown to participants during the experiment.

The two speech samples produced by the same talker in each same-talker pair (e.g. AA’) were not identical, but were extracted from two different tokens of the same CVC syllable. This ensured that the listeners were never asked to compare two identical stimulus tokens, and hence that their responses were based on speech processing rather than auditory change detection (see Andics *et al.*, 2007). As Figure 6.2 furthermore illustrates, 500ms of pink noise was played between the samples in each pair in order to minimise the influence of echoic memory (Crowder, 1982;

Witkin, 1990). A 900ms-long filler beep was presented between the two pairs in order to more explicitly separate the stimuli into pairs. The interstimulus intervals (ISIs) were consistent across trials. Between the samples in each pair and the pink noise the ISIs were 200ms long, and between the end/start of each pair and the filler beep the ISIs were 300ms long. The duration of the speech samples was on average 469ms (*SD* = 47), and the mean durations were fairly consistent across facewear conditions (control: $\bar{x}$ = 472ms, *SD* = 73; helmet: $\bar{x}$ = 454ms, *SD* = 54; tape: $\bar{x}$ = 480ms, *SD* = 81).

Finally, all intensities were normalised using the 'Scale intensity' function in *Praat*. The pink noise was presented at 50dB and the filler beep at 60dB. This provided a comfortable hearing level in relation to the speech samples, which were normalised to 70dB. Normalisation for amplitude was considered necessary, because otherwise loudness variation could have additionally influenced talker discrimination (see e.g. Miller, 1978, who showed that vowel choices can be affected by loudness variation in a timed classification task).

6.2.1.4 Procedure

Prior to taking part in the study, participants were informed about the procedure of the experiment so that they could grant their informed consent to participate. Otherwise, no detailed information about the background of the study was given to them, so as to avoid biasing their responses. For example, participants were not told that different types of facewear conditions were included in the dataset. This was intended as a way of avoiding some form of 'perceptual compensation' for facewear effects, which could be based on the listeners' experience with such coverings over a talker's mouth (e.g. through personal experience, or TV viewing). Informal interviewing after completion of the experiment indicated that some participants had suspected that the test material was electronically manipulated in some way (e.g. through digital band-pass filtering).

All participants were tested individually in a sound-attenuated IAC (Industrial Acoustics Company) booth at the MARCS Institute. The speech material was presented to them through high-quality Sennheiser HD 650 headphones. The stimuli were played back using experimental control software designed specifically for the purpose of this experiment in *Matlab*.[53]

Participants were advised that their task in each trial of the experiment was to identify the pair of speech samples in which they perceived the talker to be the same. They were instructed to make their selection by pressing one of two shift keys on a standard desktop computer keyboard, which were clearly labelled as *pair 1* (assigned to the left shift key) or *pair 2* (right shift key). In other words, when the listeners believed that the two speech samples in pair 1 had been produced by the same talker, they would press the left response key, and when they perceived the two samples in pair 2 as originating from the same talker, they would press the right key. Participants were informed that the experiment was timed (i.e., that reaction time was measured), and that no feedback about the correctness of responses would be given. On a side note, neither handedness of the listeners nor the assignment of shift keys to the response options was counterbalanced. However, this is not considered problematic because there was no statistical evidence for a response bias in the control condition (for details see Appendix D.4).

The order of trials was pseudo-randomised across listeners, a measure taken so as to compensate for practice and fatigue/boredom effects. The experiment was presented in two blocks. Between the blocks the participants took a short break, during which they had an informal conversation with the experimenter (the author). This was intended as a way of distracting them from the task. Additionally, there were four built-in self-paced breaks per block (minimum break of 10s). A within-group design was applied, whereby each participant was exposed to all 432 trials (12 pairs x 3 facewear conditions x 6 consonants x 2 presentation orders). Before the start of the experiment, the participants undertook a brief training session, during which they

[53] Thanks to Benjamin Schultz for providing the *Matlab* code, and for his assistance with the experimental design and data analysis.

were familiarised with the experimental interface and procedure. They also had the possibility to adjust the playback volume to a comfortable hearing level.

6.2.2 Results

The performance measures were response accuracy (proportion of correct talker discriminations) and response time. Response time is a sensitive indicator of the participants' performance from which cognitive processes can be inferred, particularly when ceiling effects are observed. It was measured from the offset of the second sample of pair 2, to keypress. The response accuracy and response time data were analysed separately by conducting a series of four-way repeated-measures ANOVAs using *IBM SPSS Statistics V.19.0.0.1*. The independent variables were 'facewear' (control, helmet, tape), 'consonant' (/t p s f n m/), 'order' (order 1 = same-talker pair + different-talker pair, order 2 = different-talker pair + same-talker pair), and '(different-talker) pair' (AB', AC', AD', BA', BC', BD', CA', CB', CD', DA', DB', DC').

Effects are reported as significant when $p < .05$. Where Mauchly's test indicated that the assumption of sphericity had been violated, the degrees of freedom, *p*-values and effect sizes (η_p^2) were adjusted using the Greenhouse-Geisser correction (the correction factor ε is listed in the corresponding results table in such cases). In the following sections, the results are reported in the form of averages across presentation order. For details of the effect of order see Appendix D.4. All ANOVA results tables can be found in Appendix D.5 (see Tables D.49 to D.52).

Overall, 78.2% (*SD* = 5.5) of all talker discriminations were correct. This shows that the participants on average performed considerably better than chance level (50%). There was a higher proportion of correct responses for order 1 (81.2%) than order 2 (75.1%). No gender effect was found (female listeners = 79%, male listeners = 77.2%). Individual response accuracy for the 24 listeners varied between 64.6% and 88.9%. A series of one-sample *t*-tests indicated that the overall accuracy score ($t(23)$

= 24.9, $p < .001$), as well as the scores for each participant individually (averaged across facewear and order) significantly differed from chance level (for details see Appendix D.6, Table D.53). No individual accuracy score markedly deviated from the scores of the rest of the listeners. Therefore, the data for all 24 listeners were included in the further analysis (statistical outliers were defined as those falling into the 1.5 interquartile ranges below the 25th and above the 75th percentile).

6.2.2.1 Effect of facewear

The mean percentage correct scores for all three facewear conditions (control, helmet, tape) are plotted in Figure 6.3. As the figure shows, the highest accuracy and a near-ceiling effect emerged in the control condition ($\bar{x}$ = 92.6%, *SD* = 13.4). The listeners' response accuracy overall dropped in both the helmet ($\bar{x}$ = 74.2%, *SD* =14.9) and tape ($\bar{x}$ = 67.6%, *SD* = 11.1) conditions. One-sample *t*-tests for each condition separately indicated that all three scores significantly differed from chance level (control: $t(23) = 31.7$, $p < .001$; helmet: $t(23) = 16.2$, $p < .001$; tape: $t(23) = 15.8$, $p < .001$).

The statistical analysis of the data revealed a significant main effect of facewear on response accuracy [$F(2,46) = 234.27$, $p < .001$, $\eta_p^2 = .91$]. *Post-hoc* Bonferroni-adjusted pairwise comparisons were carried out to examine whether the differences between facewear conditions shown in Figure 6.3 were significant. It was found that response accuracy in the helmet condition was significantly lower than in the baseline, and that the accuracy score in the tape condition was significantly lower than in the helmet condition (*p*s < .001).

The results so far indicate that the listeners were overall very good at determining the pair (out of two pairs presented) in which speech sample 1 and speech sample 2 originated from the same talker. The listeners' performance in this task was highest in the control condition, where all samples had been produced without the talkers wearing facewear. However, when the first sample in each of the two pairs came

from a talker whose face/mouth was obstructed either by a helmet or a piece of tape during speech production, the listeners' ability to accurately detect the same-talker pair significantly decreased. In other words, when the two samples in each pair did not match in terms of the facewear conditions they were elicited in (helmet versus control, and tape versus control), talker discrimination was made significantly more difficult.

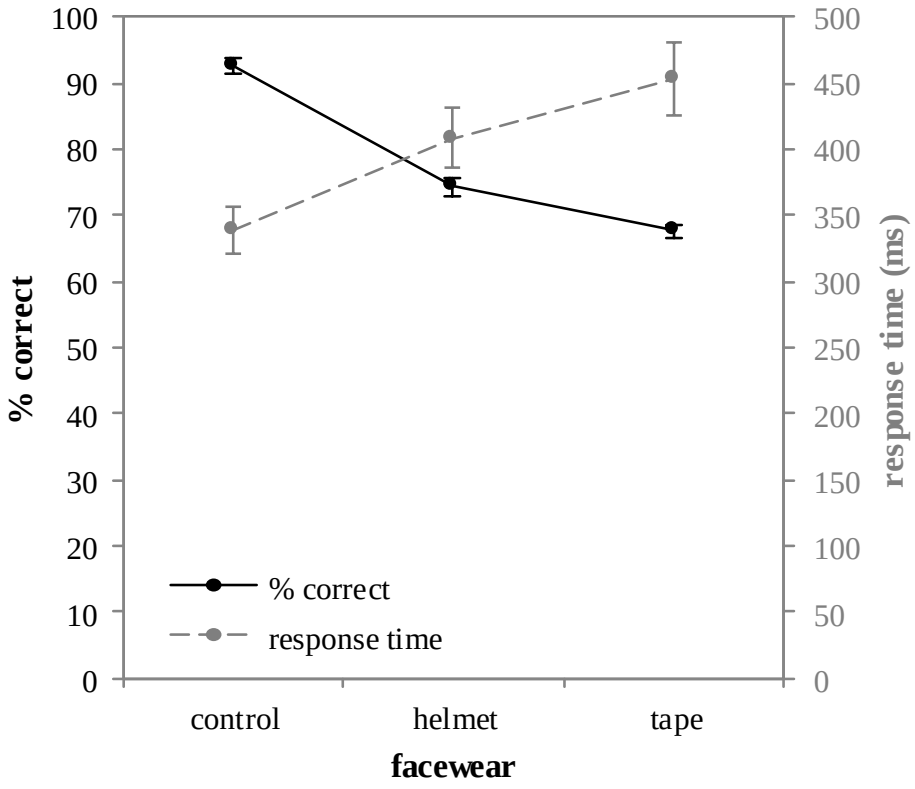

Figure 6.3. Response accuracy (see left y-axis) and response time (see right y-axis) obtained in the control, helmet, and tape conditions, averaged across listeners. Talker discrimination accuracy significantly differed for all three conditions (*p*s < .001). Response time significantly increased in the helmet and tape conditions compared to baseline (*p*s < .001), and was significantly higher in the tape than in the helmet condition ($p < .01$). The error bars show the standard error of the mean.

Next, response times were analysed. It was hypothesised that the more uncertain the listeners were about which pair of speech samples originated from the same talker, the longer it would take them to respond. To ensure ease of comparison, the mean response times for all three conditions are plotted in Figure 6.3 together with the

mean accuracy scores. Note that as the response times were similar in the correct and incorrect trials, the data were averaged across all trials. Figure 6.3 illustrates that the response times increased along with the number of errors participants made. As expected, it took the listeners longer to make their selections as the task (of choosing the same-talker pair) became more difficult. Response time was on average longest in the tape condition ($\bar{x}$ = 453ms, *SD* = 27), followed by the helmet condition ($\bar{x}$ = 408ms, *SD* = 23). The listeners responded fastest in the control condition ($\bar{x}$ = 339ms, *SD* = 18), where their performance was best (highest accuracy).

ANOVA revealed that these differences were statistically significant. There was a significant main effect of facewear on response time [$F(1,31) = 32.75$, $p < .001$, $\eta_p^2 = .59$]. *Post-hoc* Bonferroni-adjusted pairwise comparisons showed that the response times measured in all facewear conditions significantly differed from each other ($p < .001$ for control compared to helmet, $p < .001$ for control compared to tape, and $p < .01$ for helmet compared to tape).

6.2.2.2 Effect of consonant

In the previous section, the results from the talker discrimination experiment averaged across the six tested monosyllables were reported. To examine whether the participants' ability to distinguish between talkers varied with the segmental content of the speech samples, the data were subsequently split up according to the different consonant-vowel utterances. The mean percentage correct scores brought about by each of the six test syllables as a function of facewear are illustrated in Figure 6.4. A series of one-sample *t*-tests indicated that the scores obtained for all syllables in all facewear conditions significantly differed from chance level (for details see Appendix D.6, Table D.54).

The statistical analysis revealed a significant main effect of facewear on response accuracy (*p*s < .001) for all test syllables, i.e., /pɑː/ [$F(2,46) = 79.77$, $\eta_p^2 = .78$], /tɑː/ [$F(2,46) = 118.71$, $\eta_p^2 = .84$], /fɑː/ [$F(2,46) = 68.27$, $\eta_p^2 = .75$], /sɑː/ [$F(2,46) =$

40.27, η_p^2 = .64], /mɑː/ [F(2,46) = 109.04, η_p^2 = .83], and /nɑː/ [F(2,46) = 38.06, η_p^2 = .62]. This means that the facewear effects on the listeners' ability to accurately discriminate between the talkers (as reported in the previous section) occurred irrespective of the type of syllable presented to the listeners in an experimental trial. Hence, facewear seems to have changed the perceptual qualities of all consonants (+ the vowel) to an extent that talker discriminability was diminished.

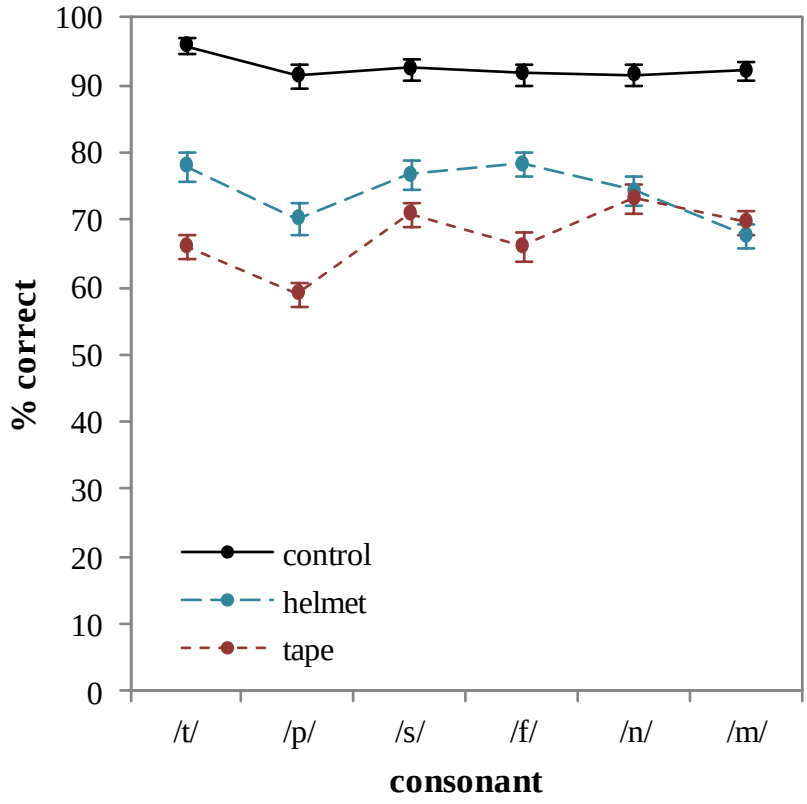

Figure 6.4. Response accuracy for all six consonants as a function of facewear. Mean accuracy scores were throughout significantly higher for control than helmet and tape, and for all consonants except /n/, /m/, and /s/ significantly higher for the helmet than the tape. The ranking of consonants (highest to lowest accuracy) differed across facewear conditions, indicating that the consonants were not equally affected by facewear type. The error bars show the standard error of the mean.

Post-hoc Bonferroni-adjusted pairwise comparisons showed that response accuracy significantly dropped from control to helmet, from control to tape, and from helmet to tape (ps < .001) for the majority of consonants tested (see Figure 6.4). The only exceptions to this pattern were /n/ (p = 1.000), /m/ (p = .714), and /s/ (p = .132), for

which the scores obtained in the helmet and those obtained in the tape condition did not significantly differ (/m/ even caused a slightly higher accuracy score in the tape condition).

As expected from the patterns shown in Figure 6.4, the statistical analysis revealed a significant main effect of consonant on talker discrimination accuracy [$F(5,115) = 9.54$, $p < .001$, $\eta_p^2 = .29$], and also a significant interaction between consonant and facewear [$F(10,230) = 6.12$, $p < .001$, $\eta_p^2 = .21$]. To explore the interaction further, ANOVAs were rerun for each level of facewear. It was found that the main effect of consonant on talker discrimination was significant in the control [$F(5,115) = 3.10$, $p < .05$, $\eta_p^2 = .12$] and helmet conditions [$F(5,115) = 7.55$, $p < .001$, $\eta_p^2 = .25$], and also in the tape condition [$F(5,115) = 8.23$, $p < .001$, $\eta_p^2 = .26$].

These results suggest that the listeners' performance in the talker discrimination task was in all three facewear conditions considerably influenced by the type of consonant embedded in the /Cɑː/ syllable. That is, the listeners' ability to detect the same-talker pair was consistently better in some trials (in which a particular consonant was presented) than in others (where another consonant was presented). As a reminder, the same consonant was presented four times within a trial. This finding suggests that the segmental content of the speech samples – here, the consonants, as the vowel /ɑː/ was kept constant – had in fact made a bigger or smaller contribution to the listeners' success in discriminating between talkers.

It was, however, also found that the 'ranking' of consonants varied across facewear conditions. This means that the *magnitude* of the reduction of response accuracy in the helmet and tape conditions compared to the control condition was dependent on (and varied with) the particular consonant embedded in the test syllable. Put differently, the consonants which brought about the highest talker discrimination scores and those which led to the lowest rates differed between conditions. The listeners' performance approximated ceiling level across the consonants in the control condition, with accuracy scores per consonant ranging from 96% for /t/ to 91.5% for /p/ (see Figure 6.4). Response accuracy was consistently lower, but generally more variable, in the helmet condition. Here, performance was best when /f/ was presented (78.3%) and worst when /m/ was presented (67.7%). In the tape

condition, /n/ scored highest (73.3%) and /p/ scored lowest (59.2%). On the whole, the ranking of consonants was /t/ > /s/ > /m/ > /f/ > /n/ > /p/ for control, /f/ > /t/ > /s/ > /n/ > /p/ > /m/ for the helmet, and /n/ > /s/ > /m/ > /t/ > /f/ > /p/ for the tape.

However, *post-hoc* Bonferroni-adjusted pairwise comparisons across all levels of consonant (within a given facewear condition) showed that these rankings should be interpreted cautiously; the differences between the percentage correct score for a particular consonant and the scores for each of the other consonants were not always significant. In the control condition, the highest response accuracy for /t/ significantly differed only from the percentage correct scores for /p/, /n/, and /f/ (*p*s < .05); there were no other significantly different consonant pairs. In the helmet condition, /f/ significantly differed from /m/ ($p < .001$) and /p/ ($p < .01$), /t/ significantly differed from /m/ ($p < .01$) and /p/ ($p < .05$), and /s/ and /n/ each differed from /m/ (*p*s < .05). For tape, /n/ and /s/ significantly differed from /p/ (*p*s < .001), and /m/ from /p/ ($p < .01$).[54]

Finally, no significant effects of consonant on the response time measures were found in any of the facewear conditions (*p*s > .05).

6.2.2.3 Effect of pair

The previous sections have shown that talker discriminability was reduced when facewear speech was involved in the task (effect of facewear), and that some consonants supported the listeners' ability to distinguish between unfamiliar talkers more than others (effect of consonant). Moreover, talker discrimination rates brought about by certain consonants – and hence the perceptual properties of the consonants

[54] The finding that response accuracy for each consonant was reduced to a varying degree by the helmet and tape was additionally confirmed by computing Pearson correlation coefficients between the mean percentage correct score per consonant for the baseline and the corresponding scores for each of the facewear conditions, as well as between the scores for the two facewear conditions (*N*s = 6). There were no significant correlations between control and helmet (Pearson's $r = .456$, $p = .364$), control and tape (Pearson's $r = –.022$, $p = .967$), or helmet and tape (Pearson's $r = .148$, $p = .779$).

– were found to be differently affected by facewear (interaction facewear x consonant). This section explores whether the observed patterns occur consistently for all talkers, or whether the changes to the perceptual qualities of the consonants caused by facewear are more likely to be talker-specific.

A closer inspection of the data suggests a highly complex relationship between the type of consonant that was presented to the listeners, the facewear condition that the speech was produced under, and the specific combination of talkers in a particular experimental trial. To help understand the complicated patterns that arose in the present data, Figure 6.5 shows the mean accuracy scores (for each facewear condition) obtained for all 12 different-talker pairs averaged across consonants.[55] The pairs are ordered on the x-axis (from high to low) according to the percentage correct score averaged across all trials in which a particular talker was the *target*. For example, talker D was the target in trials consisting of the different-talker pairs DA', DB', and DC' (all of which were followed or preceded by the same-talker pair DD'), but *not* in trials consisting of the different-talker pairs AD', BD', and CD' (which were followed or preceded by the same-talker pairs AA', BB', and CC', respectively).

On the whole, talker D was most successfully discriminated (88.3%), followed by talker B (81.4%), talker C (71.5 %), and talker A (71.3%). Dividing the data by type of facewear showed that talker D obtained 93.5% correct talker discriminations in the control, 83.7% in the helmet, and 87.8% in the tape condition. For talker B, the response accuracy dropped from 89.7% in the control to 79.2% in the helmet and 75.5% in the tape condition. Talker C scored 93.2% in the control, 69% in the helmet, and 52.4% in the tape condition. Lastly, performance for talker A declined from 94.1% (control) to 65.2% (helmet) and 54.6% (tape). This indicates that facewear seems to have affected the speech of some talkers more than of others (overall more strongly for talkers A and C than for talkers B and D).

[55] Given that the listeners had a 50% chance of selecting the correct (i.e., same-talker) pair, it is not necessary to include the same-talker pairs in the figure.

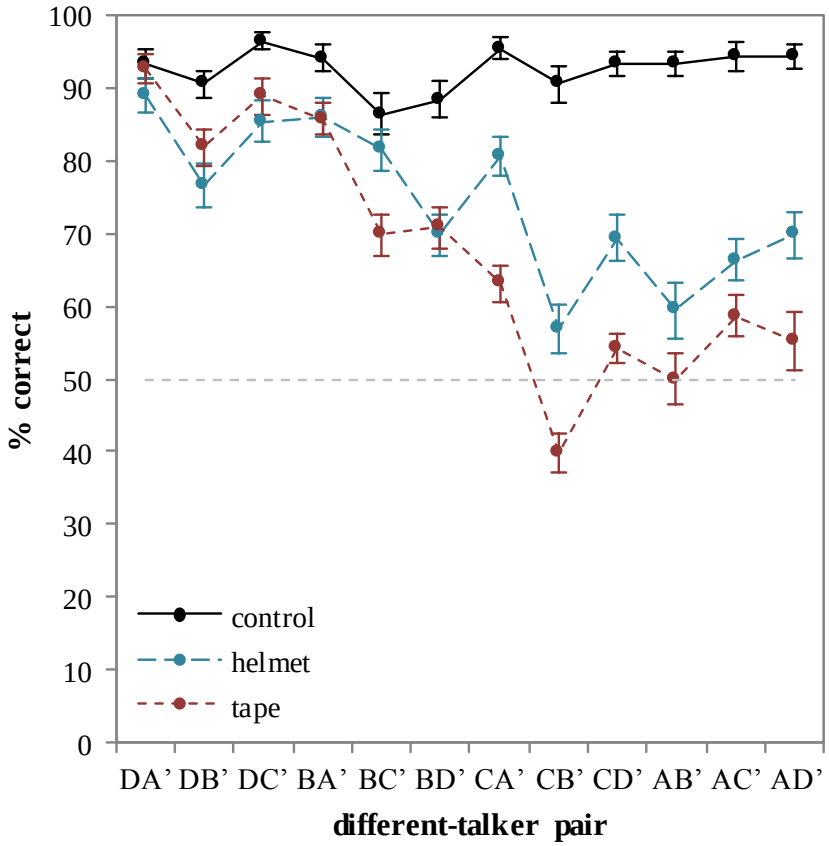

Figure 6.5. Response accuracy for all twelve different-talker pairs as a function of facewear (averaged across consonants). The pairs are ordered along the x-axis according to the percentage correct score averaged across all trials in which a talker was the target. The dashed horizontal line represents chance level (50%). The error bars show the standard error of the mean.

Figure 6.5 demonstrates that the differences in response accuracy obtained in the control condition and that brought about in the helmet and tape conditions, respectively, were smaller for some talkers than for others. By and large, the facewear effects were less pronounced for talker B, and especially for talker D. This means that the speech samples produced by these two talkers could still be quite well distinguished from those of other talkers even when talkers B and D were wearing facewear while producing the speech stimuli. By contrast, talkers C and A were considerably less distinguishable from other talkers when they were speaking through facewear (sometimes performance was even below chance level). The variation across pairs was confirmed by a significant main effect of pair on response

accuracy [$F(6,137) = 35.75$, $p < .001$, $\eta_p^2 = .61$], and also a significant interaction between pair and facewear [$F(10,230) = 16.56$, $p < .001$, $\eta_p^2 = .42$].

Furthermore, the 'unequal' detrimental effect of facewear on talker discriminability occurred across the tested consonants. This was confirmed by a significant three-way interaction between pair, facewear, and consonant [$F(110,2530) = 2.81$, $p < .001$, $\eta_p^2 = .11$]. Illustrations of the mean accuracy scores for each different-talker pair, facewear condition, and consonant separately can be consulted in Appendix D.7 (see Figures D.2 to D.7). To further assess this interaction, Pearson correlation coefficients were computed between the mean accuracy score per pair in the control and each of the facewear conditions, and between the scores for the two facewear conditions. This was done for each consonant separately (see Appendix D.8).

Only a small proportion of the correlations were significant. These were the comparisons between the helmet and tape conditions for /m/ ($p < .001$), /n/, /f/ (ps $< .01$), and /t/ ($p < .05$), as well as between the control and helmet conditions for /n/ ($p < .05$). This suggests that the negative effects of the tape and/or the helmet on talker discrimination based on these consonants were relatively stable across talkers.

Most correlations, however, were found to be non-significant. This suggests that the facewear-induced changes to the perceptual qualities of most consonants occurred in a talker-specific manner. For example, the listeners' ability in /t/ trials to tell apart talker A from talker B was significantly reduced when talker A spoke through the helmet, whereas the discrimination of talker D from talker B was less (or not at all) affected when talker D spoke through the helmet. In this scenario, talker A's speech was more affected by the helmet than talker D's speech.

6.2.3 Phonetic cues to talker discrimination

The results presented in the previous sections showed that facewear minimised the listeners' success in discriminating between unfamiliar talkers, and that some consonants provided more cues to successful talker discrimination than others. The extent to which facewear impacted on the listeners' performance in the task was dependent on the type of consonant embedded in the tested /Cɑː/ syllables. The degree of interference of facewear with talker discrimination also varied across talkers. On the basis of these observations we can argue that the amount of idiosyncratic information that a specific segment carries – i.e., information which facilitates decisions about whether or not two speech samples originate from the same talker – will change (in a talker-specific manner) when the segment is produced through the helmet and tape.

The following sections present the findings from an auditory-perceptual analysis of the speech material. The goal was to relate the experimental results to the phonetic properties of the consonants that were affected by facewear (and thus led to discrimination difficulty). The analysis also meant to provide further insights into facewear effects on individual talkers' voices and speech patterns, which will be of value to the thesis more broadly.

6.2.3.1 Consonants

The highest response accuracy was obtained in the control condition for trials in which the test syllables contained the alveolar plosive /t/. The mean score for /t/ was significantly higher than the scores for /p/, /n/, and /f/. The listeners' performance in /t/ trials significantly decreased in the helmet and tape conditions, and there was also a significant drop from helmet to tape. In the helmet condition, the listeners still performed significantly better for /t/ than for /m/ and /p/. In the tape condition, /t/ scored at the lower end (relative to the other consonants).

The analysis of the data by means of Pearson correlations indicated that there is a considerable amount of idiosyncratic variation in the extent to which talker discrimination based on /t/ differed when /t/ had been produced through the helmet or tape. There was no significant relationship between the scores obtained in the control condition and those in the helmet and tape conditions, respectively. This suggests that the perceptual – and hence the acoustic and possibly articulatory – properties of different talkers' /t/ productions were not equally affected by facewear.

Auditory analysis accompanied by visual inspection of spectrograms of the /t/ data in *Praat* confirmed that the characteristics of the plosive burst varied across talkers. This accords with the literature, which reports that the production of alveolar stops differs across individuals, e.g. with regard to the location and duration of the closure, or the strength, speed, and direction of the closure release (Foulkes *et al.*, 2010). Here, most talkers (A–C) produced single transients during the release of the closure, while one talker (D) produced multiple transients and a high-energy burst suggestive of a weak ejective. Voice onset time (VOT) was similar across talkers, but still revealed some inter-talker variation (e.g. it was shorter for C than for other talkers). The burst intensity relative to the vowel intensity also differed across talkers. For example, talker B produced high-energy frication at closure release, while the intensity of the bursts produced by talkers A and C was quite low.

Most notably, the intensity of the transients remained comparatively unchanged for some talkers (A+B), but was reduced (C) or enhanced (D) for others in the tape condition. In the helmet condition, the burst intensity was reduced for all talkers. The latter conforms to the findings from the acoustic study of voiceless plosives (Chapter 4), where it was observed that the burst intensity was considerably lower in the helmet than in the control condition.

Moving on to the trials in the talker discrimination experiment in which the bilabial plosive /p/ was presented, one finds that the listeners' performance for /p/ was overall rather low. The proportion of correct talker discriminations based on /p/ declined when /p/ was produced through the helmet and, even more so, through the tape. Specifically, /p/ scored significantly lower than /f/ and /t/ in the helmet condition, and lower than /m/, /n/, and /s/ in the tape condition.

The lower performance for /p/ suggests that /p/ overall carried less talker-discriminating information than did /t/. This observation is in accordance with previous research which reports that bilabial stops are in general less distinct than stops produced at other places of articulation (Kewley-Port, 1983; Hawkins & Stevens, 1987). Nevertheless, auditory analysis revealed some degree of inter-talker variation in the control data. Namely, VOT varied across talkers (it was shorter for C, and longer for D), as did the burst intensity relative to vowel intensity in the control condition. One talker (A) produced low-energy bursts and only marginally visible transients, while others (B+C) generated clearly visible transients. Also, talker D again produced a high-energy, ejective-like stop consonant.

Taken together, auditory-perceptual analysis gave the impression that /p/ was quite strongly affected when the talkers' faces were disguised, and that /p/ was generally more vulnerable to acoustic modifications caused by facewear than was /t/. This may in part be explained by the fact that the natural movement of the talkers' lips – an intrinsic requirement for the production of bilabial sounds – was considerably constrained when the mouth was occluded (and especially when it was taped shut). In addition, the energy of the burst and transient decreased for most talkers (A–C) in the helmet condition, and for some talkers (A+D) in the tape condition, while for one talker (B) the release transients notably increased in the tape condition. This again confirms the results of the acoustic study discussed in Chapter 4, where the burst intensity of /p/ showed a tendency to decrease when the plosive was produced through the helmet and tape.

Next, the listeners' performance in trials involving the alveolar fricative /s/ was equally high in all facewear conditions, and the performance for /s/ was not significantly different from that for /t/. The listeners' response accuracy obtained for /s/ dropped in the helmet and tape conditions in comparison to the baseline, but was in both conditions still high compared to the rest of the consonants (significantly higher than /m/ and /n/ in the helmet condition, and significantly higher than /p/ in the tape condition).

The overall high performance in /s/ trials may be partly connected with the generally large amount of acoustic energy in /s/. This may have kept the perceptual effects of

facewear-induced sound energy absorption and acoustic filtering (as reported in §4.3) within limits. Moreover, the 'normal' articulation of /s/ could be fairly well sustained even when the fricative was produced through the tape. As mentioned earlier, the tape (and also the helmet to some extent) notably impaired natural lip motion during speech production. However, as lip motion is less critical for the production of alveolar sounds like /s/ (and /t/) than for the production of, say, bilabial and labiodental consonants, the articulatory constraints imposed by this particular sort of facewear were not problematic in the case of /s/. This may explain why the response accuracy for /s/ produced through the tape did not significantly differ from the response accuracy for /s/ obtained for the helmet speech.

The auditory analysis of /s/ again revealed a high amount of variation across talkers. In particular, the frication noise produced by talker D showed considerably higher intensity than the intensity of the frication noise in all other talkers' /s/ productions. However, when /s/ was spoken through the helmet or tape, the intensity of the turbulent airflow relative to the intensity of the vowel decreased for most talkers. The reduction of acoustic energy in /s/ in the facewear conditions is in keeping with the findings from the acoustic study of voiceless fricatives (Chapter 4), in which the intensity of /s/ was found to be significantly lower in the helmet and tape compared to the control measures (ps < .001). The intensity variation across talkers in the tape condition may in part be ascribed to how firmly the tape was adhered to the talkers' mouth/cheeks. As this slightly differed between talkers (despite best efforts to ensure consistency across talkers), the channel for the air to escape from the vocal tract at the side of the tape may have been wider or narrower, and might therefore have caused more or less additional turbulences.

By contrast with /s/, the proportion of correct talker discriminations in experimental trials involving the labiodental fricative /f/ was lower in the control condition (but still close to ceiling). The high performance may again be ascribed to the fairly large amount of between-talker variation, whereby /f/ was characterised by overall low intensity for some talkers (A+C) but comparatively high energy for others (B+D).

There was again a drop in performance in the helmet and additionally in the tape condition. The intensity of the frication noise tended to decrease for most talkers (A–

C) in both facewear conditions. The reduction of intensity when /f/ was produced through the helmet coincides with the significant intensity drop observed in the acoustic study ($p < .001$). However, it was found that the intensity of /f/ was higher in the tape than in the helmet condition for most talkers (B–D). As with /s/, this could be the consequence of amplification of the frication noise and/or additional turbulences caused by the tape acting as a secondary constriction in front of the talker's mouth.

Interestingly, /f/ scored higher in the helmet condition (significantly higher than /m/ and /p/) than in the tape condition. This difference between the response accuracy obtained for /f/ spoken through the helmet and the (lower) accuracy obtained in the tape condition was fairly consistent across talkers (as indicated by significant Pearson correlations). Once again, one possible explanation for the reduced proportion of correct talker discriminations in the tape condition is that the movement of the lower lip, which is necessary for the production of labiodental sounds, was more strongly perturbed by the tape than by the helmet.

Up to this point, the participants' performance in relation to the articulatory and acoustic variation observed across talkers has been discussed for the trials in which the four oral consonants /t/, /p/, /s/, and /f/ were presented to listeners. By and large, the listeners' ability to distinguish between talkers dropped when the consonants were produced through the helmet, and even more so when they had been spoken through the tape (with very few exceptions). The results for the two nasal consonants tested in this study exhibited a different pattern altogether.

In accordance with the oral sounds, the mean accuracy scores obtained for /m/ and /n/ decreased in both facewear conditions. In the helmet condition, both /m/ and /n/ (together with /p/) scored at the lower end (/m/ significantly lower than /t/, /s/, /f/, and /n/; /n/ significantly lower than /s/). However, significant Pearson correlations indicated that the proportion of correct talker discriminations obtained for the helmet and tape conditions, respectively, did not significantly differ between talkers (/m/ spoken through the tape even scored slightly higher on average than /m/ spoken through the helmet). That is, the perceptual qualities of the two nasals were overall less affected by the tape than were the perceptual properties of the oral consonants.

This finding might have been expected from the fact that the production of nasal consonants could be reasonably well maintained in spite of the facewear covering the talkers' mouth. Each talker's nose was completely unconcealed in the tape condition, and was only partly occluded in case of the helmet, for which reason the air could still escape unhindered through the nostrils.

Perceptually, /m/ and /n/ produced through facewear on occasion differed from the same sounds produced in the control condition. For most talkers, the nasals produced through the helmet gave the auditory impression of denasality. This may have been triggered by acoustic absorption of nasal formants caused by the sound-absorbing outer shell of the helmet that was concealing each talker's nose. In the tape condition, the perceptual quality of /m/ changed for one talker (C) consistently to a velar nasal [ŋ], which may be indicative of articulatory compensation. For another talker (A), /m/ mostly sounded like a labiodental approximant [ʋ]. This may have been the result of the tape preventing the lips from forming a complete bilabial closure, instead only permitting an approximation of the lower lip to the upper teeth (thus leaving air to escape from the side of the tape). The misperception of /m/ as [ʋ] reflects the results from the consonant identification experiment presented in Chapter 5 (quiet listening condition). Here, 37 out of 86 /m/ presentations in the auditory-only condition were wrongly identified as /v/ (see Appendix D.1, Table D.9).

6.2.3.2 Vowel

Even though the focus of the current study lies on the talker-discriminating power of consonants spoken through facewear, it was considered worthwhile to take a closer look at the acoustics of /ɑː/ and its possible contribution to talker discriminability. Indeed, Andics *et al.* (2007) and Bricker & Pruzansky (1966) report that it was vowel changes that made the biggest difference to talker discrimination in an experiment based on CVC words. This implies that vowels carry more paralinguistic information to assist the listeners in distinguishing between talkers than consonants do. While a comparative analysis of consonants and vowels is not possible here (the vowel was

always /ɑː/), the acoustic analysis of the first three formants of /ɑː/ nevertheless seemed worth pursuing, not least because forensic phoneticians commonly consider vowel formants to be a helpful speaker discriminant (e.g. Nolan & Grigoras, 2005; McDougall & Nolan, 2007).

Formants are acoustic resonances of the vocal tract. They are determined by the length and configuration of cavities of the supralaryngeal vocal tract, and are acoustically identified as intensity peaks in the frequency spectrum. Here, the first, second, and third formant (henceforth F1, F2, and F3) were measured automatically by means of a *Praat* script (Burg method; pre-emphasis from 50Hz; maximum formant = 5kHz; Gaussian window length = 25ms; maximum number of formants manually adjusted to 4, 5, or 6 to increase formant tracking accuracy). Measurements were taken from the steady-state portions around the temporal midpoint of each vowel (mean duration of the analysed segments was 143ms, *SD* = 75). The resultant formant values were hand-corrected, where necessary, by consulting spectrograms.

The outcome of the formant analysis is shown in Figure 6.6 (for F2 x F1) and Figure 6.7 (for F3 x F2). The figures show the means of F1, F2, and F3 (in Hertz) of /ɑː/ produced by all four talkers in each facewear condition. The corresponding figures that show all individual data points (*N*s = 12 per talker and facewear condition) are Figures D.8 to D.13 in Appendix D.7. Despite the practical limitations of this analysis (mainly concerning the small sample size), several interesting trends with respect to the effects of facewear on the first three formants of /ɑː/ can be recorded.

The most prominent changes to the formants can be observed in the tape condition. Here, the mean F1 values of all talkers' vowel productions were considerably lower than in the baseline condition, with formant shifts between 95Hz and 170Hz. By contrast, F1 remained fairly stable in the helmet condition (in-/decrease of less than 20Hz). The F1 shift in the tape condition may in part be explained by the restricted jaw movement when the talker's mouth was taped shut. Open vowels such as /ɑː/ are produced with a lowered jaw (and low tongue position). Jaw opening is associated with a high F1 (see e.g. Clark *et al.*, 2007: 290). When facewear hinders the lowering

of the jaw during speech production, F1 is quite likely to be reduced relative to F1 encountered in unperturbed speech.[56]

The averaged F2 values of all talkers' /ɑː/ productions were also lower (by 35–75Hz) when the vowel was spoken through the tape, but to a much lesser extent than F1. In helmet speech, the F2 changes were altogether more variable, with some talkers lowering F2 (by 40–70Hz) and others raising F2 (by 15–45Hz).

Lastly, the mean frequency of the third formant dropped very considerably (up to 400Hz) for all talkers when their mouths were taped closed while speaking. The helmet once again did not markedly affect F3 (except for the increase of ~130Hz for talker D).

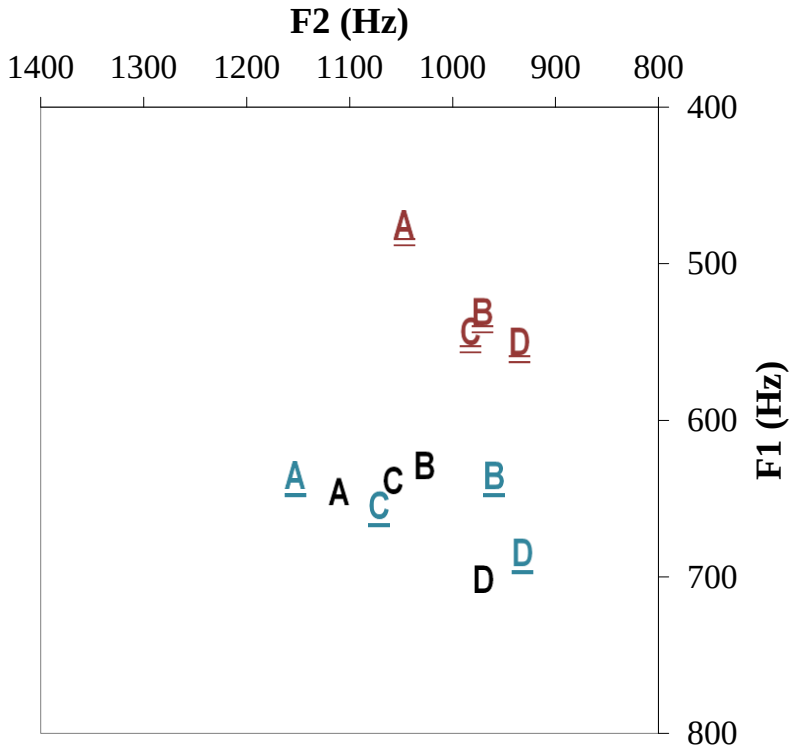

Figure 6.6. Mean first formants (F1) and second formants (F2) of /ɑː/ (in Hz) produced by talkers A, B, C, and D in the control (black/no underlining), helmet (blue/single underlining), and tape (red/double underlining) conditions.

[56] A similar argument was put forward by Bond *et al.* (1989), who ascribed the first formant changes of speech produced through an oxygen mask to the restriction of jaw movement caused by the mask. The role of the jaw was also emphasised by Abeysekera & Shahnavaz (1987), who attributed the reduced speech intelligibility caused by respirator masks to, among other things, a limited freedom of jaw motion (which is partly dependent on the respirator weight). Both studies were introduced in §2.3.3.

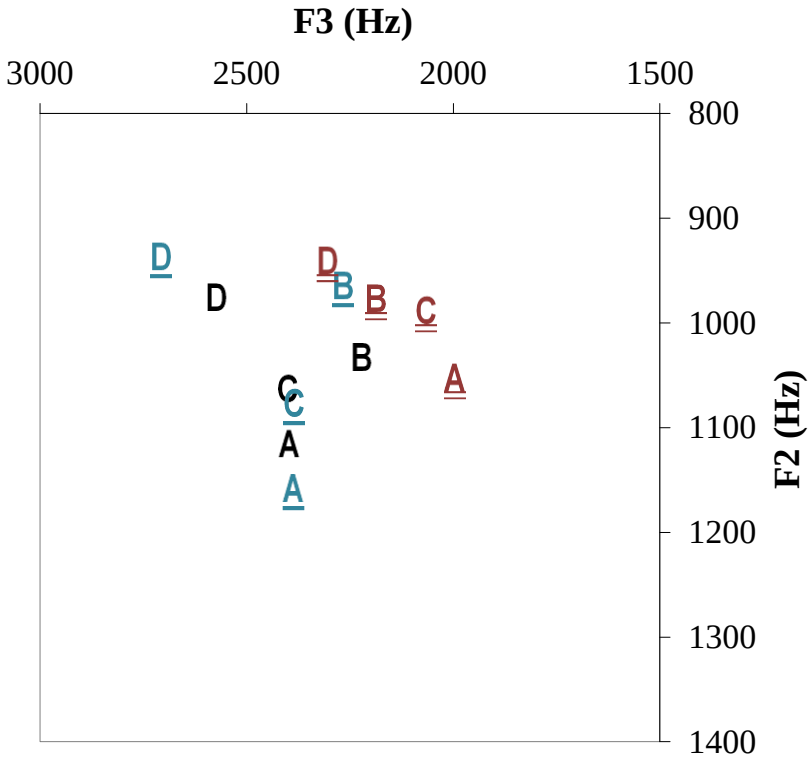

Figure 6.7. Mean second formants (F2) and third formants (F3) of /ɑː/ (in Hz) produced by talkers A, B, C, and D in the control (black/no underlining), helmet (blue/single underlining), and tape (red/double underlining) conditions.

Regarding the differences between the formant values measured for different talkers, a closer inspection of the data obtained in the control condition revealed that the formants of talker D's /ɑː/ productions on the whole differed from the rest of the talkers' /ɑː/ formants: F1 was 60–70Hz higher and F2 was 60–140Hz lower (as expected, F3 also differed, and was 170–350Hz higher). The formant data of talkers A, B, and C, on the other hand, substantially overlapped in the control condition.

Interestingly, these patterns did not hold out in facewear speech. The mean formant values of some talkers were now more similar to each other, and those of other talkers were more distinct. In the helmet condition, for example, the formant differences between talker D and the remaining talkers were reduced, with the consequence that talker B and talker D were now more similar to one another in terms of their mean F1 and F2; talkers A and C were now more distinct from talker B (in particular with regard to F2). In the tape condition, F3 produced by talker D decreased to the extent that his F3 values now overlapped more markedly with F3 for

all other talkers (especially talker B). Moreover, as the mean F1 for talker A was more strongly reduced than the mean F1 for talkers B–D in the tape speech, talker A was now most distinct from the rest of the talkers regarding F1; F1 for talkers B–D were relatively less distinguishable.

6.2.3.3 Suprasegmentals

The availability of suprasegmental cues for distinguishing between the talkers was greatly limited in the present study. Suprasegmental features are those which extend over individual segments, such as prosodic cues (e.g. stress, rhythm, loudness) and voice quality. Here, the talkers' mean fundamental frequency (F0), F0 contours, and voice quality were assessed as potential talker-discriminating suprasegmentals.

Considering F0, the only voiced segments from which F0 could be obtained were the talkers' vowel productions (with the exception of the nasals). Hence, the F0 measurements arguably reflect 'intrinsic F0' more than they reflect the talkers' overall F0 while speaking.[57]

Here, F0 was measured in *Praat* using the autocorrelation method (frame duration = 10ms; pitch floor = 75Hz; pitch ceiling = 600Hz; note that pitch analysis in *Praat* corresponds to acoustic periodicity detection). Figure 6.8 shows the F0 means and standard deviations of /ɑ:/ produced by all talkers in all facewear conditions.

Statistical analysis (by means of a two-way ANOVA with 'talker' and 'facewear' as independent variables) revealed that the main effects of talker and facewear on F0, as well as the interaction between talker and facewear, were highly significant. A series

[57] Intrinsic F0 (also termed 'intrinsic pitch') refers to the phenomenon that the mean F0 of vowels is correlated with vowel height (high vowels, like /i/ and /u/, tend to have higher F0 than low vowels such as /a/), and dependent on the voicing characteristics of obstruents in prevocalic position (see e.g. Hombert *et al.*, 1979; Shadle, 1985; Whalen & Levitt, 1995). The use of the term in the present context seems justified because the phonetic environments in which the vowels were produced (i.e., the consonants preceding and following the vowel) were the same for all talkers in all experimental conditions.

of one-way ANOVAs subsequently showed that the effect of talker was significant in all three facewear conditions (*p*s < .001), and that the effect of facewear was significant for talkers A–C (*p*s < .001), but not for talker D ($p = .258$). These results indicate, firstly, that F0 of /ɑː/ significantly differed between talkers; secondly, that F0 was significantly affected when the vowel was produced through the helmet and the tape; and thirdly, that facewear affected F0 of /ɑː/ differently for each talker. The specific patterns were revealed by *post-hoc* Tukey HSD tests.

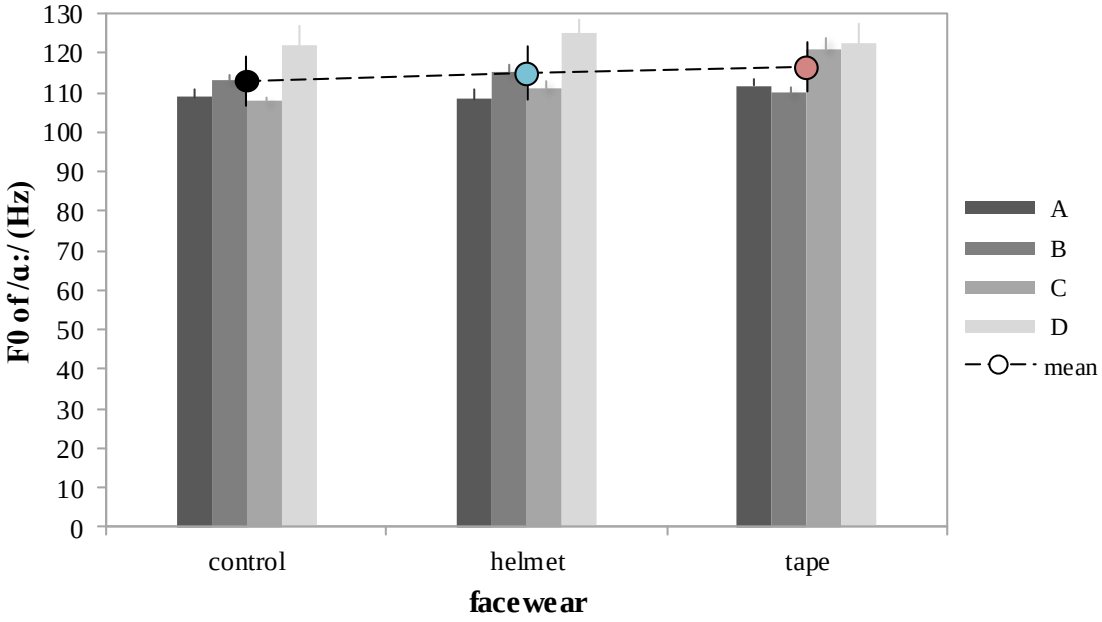

Figure 6.8. Mean F0 of /ɑː/ (in Hz) produced by talkers A, B, C, and D in the control, helmet, and tape conditions.

As Figure 6.8 demonstrates, talker D's mean F0 was significantly higher throughout (*p*s < .001) than the mean F0 for all other talkers (except talker C in the tape condition). There was, moreover, a slight tendency for talkers to increase the F0 of /ɑː/ when talking through facewear. In the tape condition, F0 was significantly higher (relative to the control condition) for all talkers except talker D (talkers A and B, *p*s < .01; talker C, $p < .001$). In the helmet condition, only two talkers significantly increased F0 (talker B, $p < .05$; talker C, $p < .01$).

As a consequence of this observed variation across talkers, two talkers would sometimes become more similar to each other and at other times more distinct in

terms of F0 of /ɑː/. For example, talkers C and D significantly differed from each other with respect to F0 in the control condition ($p < .001$), but in the tape condition this difference was non-significant. The cause of this effect was the strong F0 increase observed for talker C. This in turn produced the additional effect that talkers A and C became significantly more different regarding F0 ($p < .001$). Furthermore, the F0 reduction observed for talker B when talking through the tape resulted in talkers B and A no longer being different from each other with respect to F0 (control, $p < .01$; tape, $p = .120$). These results generally reflect the auditory impression, which confirmed the higher pitch for talker D and the prominent pitch increase for talker A in the tape condition. In particular talker A gave reason to believe that the pitch increase may on some occasions be symptomatic of increased vocal effort on the part of the talkers.

Next, F0 contours across the test syllables were assessed (see Figure 6.9). It was anticipated that varying intonation patterns produced by different talkers may have given the listeners a hint as to which speech samples were produced by the same talker. For this purpose, the data were firstly auditorily analysed. The F0 contours for all 12 test syllables (6 syllables x 2 repetitions) spoken by each talker in each facewear condition were then plotted in *Praat*. This was done after concatenating the audio files of each talker's syllable productions per facewear condition (which explains why there are no gaps between the corresponding F0 contours shown in Figure 6.9).

Neither the auditory analysis of the data nor the visual inspection of the F0 contours shown in Figure 6.9 revealed major differences between the talkers. It is therefore unlikely that the listeners used F0 variations as a cue to distinguish between the talkers. The lack of a difference in the intonation patterns between talkers can be explained by the fact that the test syllables were originally elicited in the same syntactic, semantic and phonetic environment (*He said [stimulus]*.). Also, the speaking style of all talkers was generally rather monotone, and many produced a 'list intonation'. List intonation is commonly observed when talkers have to read a large set of stimuli words or sentences for experimental purposes. The natural pronunciation of words and intonation patterns can be obscured in such cases, because the last word in the list is typically spoken with a lower pitch than earlier

words in the list (Ladefoged, 2003). All talkers in the study predominantly produced a falling intonation across conditions, which means that F0 decreased from the onset of the vowel to the offset (the only exception was the falling-rising intonation of /sɑː/ produced by talker D in the control condition). Note that the F0 contours in Figure 6.9 look different in case of the nasals (rising-falling intonation); /m/ and /n/ were the only voiced consonants in the dataset.

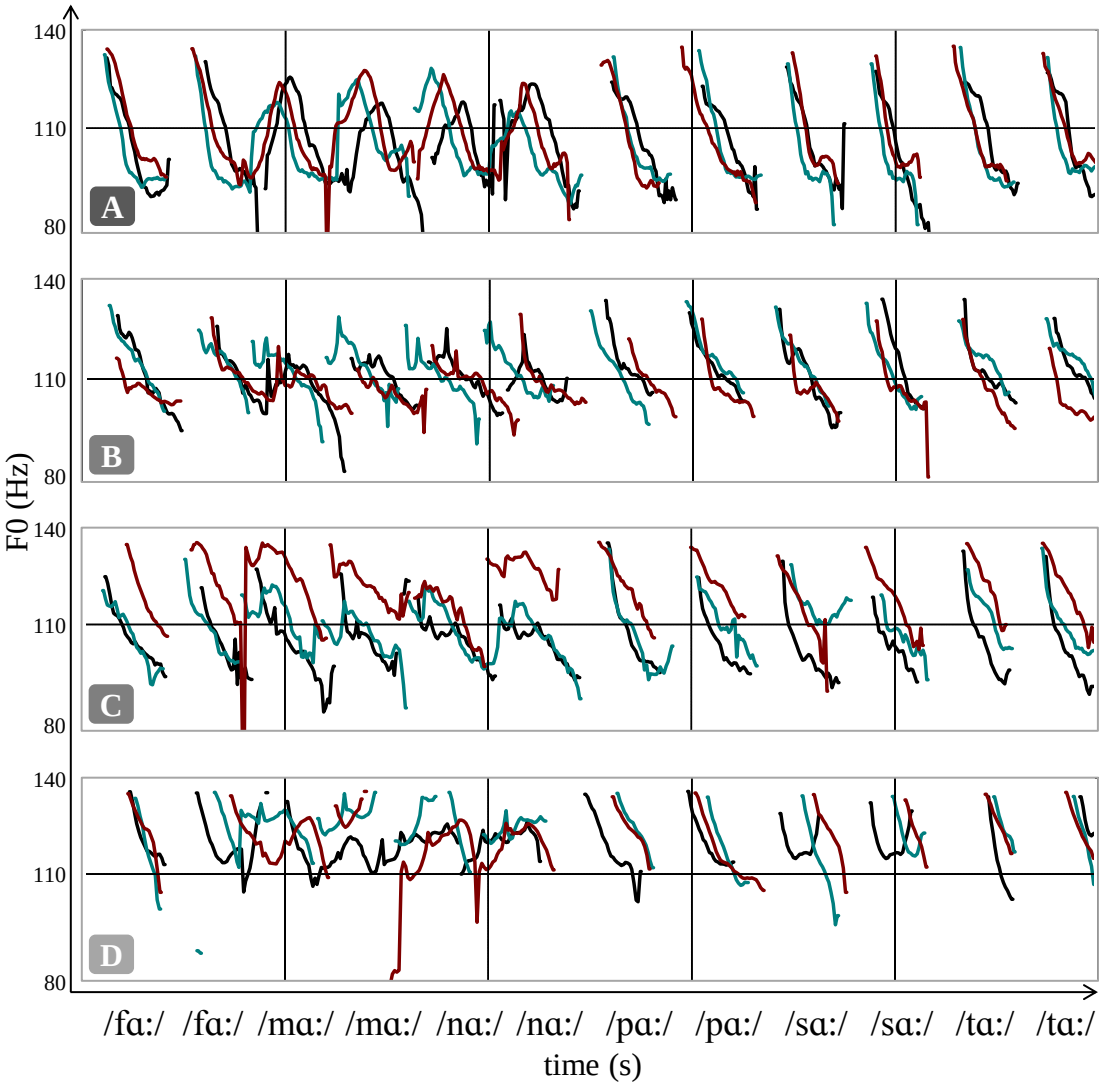

Figure 6.9. F0 contours (in Hz) for the tested CV syllables produced by talkers A, B, C, and D in the control (black), helmet (blue), and tape (red) conditions.

Finally, another suprasegmental parameter that the listeners in the experiment may have used to distinguish between the talkers was voice quality. To get an impression of how much voice quality information can be extracted from short CV syllables, a forensic phonetic expert, who is trained in using the Edinburgh Vocal Profile Analysis (VPA) protocol (Laver, 1980; French & Stevens, 2013), was asked to evaluate the speech samples with respect to phonation features (e.g. creaky voice, tremor) and vocal tract settings (e.g. nasalisation, pharyngeal constriction). Owing to the brevity of the samples, this was done in a relatively informal manner by auditorily inspecting the samples and noting down the most conspicuous findings.

The analyst reported that the CV syllables still provide an adequate amount of voice quality information in all three facewear conditions. Hence, the listeners may indeed have used voice quality as an additional indexical cue to discriminate between the talkers. In the control condition, the voice qualities of talkers A–C were overall quite similar to each other (whispery, nasalised, at times breathy and/or creaky). On some occasions, however, the talkers' voice qualities deviated from their usual properties. The listeners may have used this voice quality change as an indication that two speech samples originated from different talkers. For example, talker B could sometimes be characterised by his use of a laxer articulatory setting, and a dentalised setting associated with a fronted tongue body position (most noticeable in his dental pronunciation of [t]), and talker A by speaking with an expanded pharynx (which might explain his prominent F2 in Figure 6.6). Interestingly, talker D once again deviated most conspicuously from the rest of the talkers, in that he was auditorily most distinct on several dimensions (in addition to the higher F0, F1 and F3 and lower F2 measures discussed earlier). Talker D's voice quality indicated an articulatory setting characterised by a raised larynx, a velarised tongue setting, and quite substantial (supra)laryngeal tension. The consonants /t/ and /s/ were produced with quite 'bright' qualities corresponding to a concentration of energy at higher frequencies.

In the helmet and tape conditions, the talkers' voice qualities at times had a greater amount of whisper (glottal leakage), but the voice quality in the helmet guise was overall more modal than in the control condition (possibly because the talkers were using greater vocal effort). In addition, the speech of most talkers was more

nasalised. As noted in §6.2.3.1, the quality of the nasals was often ambiguous in the helmet and tape conditions, i.e., the place of articulation of [m] and [n] could easily be misjudged. Lastly, the impression of larynx raising in talker D's tape speech was more marked still, suggesting an even greater level of vocal effort on his part.

To sum up, it was difficult to derive any clear patterns from the voice quality judgements. However, the observations made by the expert listener once more confirm those made for the other parameters discussed in this chapter (vowel formants, F0, etc.), namely that in the helmet and tape conditions, two talkers would sometimes become more similar to each other and at other times more distinct in terms of their voice qualities and vocal tract settings.

6.3 General discussion of Experiment 5

The final part of this chapter discusses the results from Experiment 5 in view of the preceding research on the topic (see §6.1) and of the findings from the acoustic and perceptual studies presented in Chapters 4 and 5. The experiment investigated the ability of lay listeners to distinguish between unfamiliar talkers when all the listeners have available for comparison are short consonant-vowel syllables. The two main goals of this work were to gain further insights into the extent to which the segmental content of speech has an impact on talker discrimination, and to discover how two types of facial disguise affect the listeners' performance in this task.

The listeners' assignment in each trial of the experiment was to decide which pair of speech samples (out of two pairs presented) had been produced by the same talker ('two-interval forced-choice' procedure). That is, they were to determine whether two samples of speech were produced by the same talker or by two different individuals. The speech samples were very short (~500ms on average) and highly controlled (e.g. for presentation volume, interstimulus intervals, presentation order of trials, or presentation order of same-talker and different-talker pairs within a trial). The consonants varied between /t/, /p/, /s/, /f/, /n/, and /m/; within a particular experimental trial the same consonant was presented. The syllable nucleus was consistently the open back vowel /ɑː/. Speech data from four of the male talkers recorded for the AVFC corpus (see Chapter 3) were chosen for testing.

The 24 participants in this study evaluated the talkers' voices and speech productions in a control condition, where all speech samples originated from talkers whose faces were undisguised. Owing to the nature of the experimental design and the resultantly large number of test trials per condition, the listeners' performance was assessed in only two facewear conditions: motorcycle helmet and piece of tape across the mouth/cheeks. This choice of facewear was motivated by the results from the studies presented in previous chapters, and the author's casework experience (see §1.1.2.1).

In total, 78.2% of all talker discriminations were found to be correct. This suggests that the listeners were able to correctly distinguish between the talkers significantly better than chance level (50%). Considering how limited the speech material

presented to the listeners was, and (comparatively speaking) how little indexical information could have been encoded in the stimuli, this result once again demonstrates the remarkable ability of the human speech perception system to extract talker-related information even from single consonants and vowels.

To arrive at a conclusion about whether two speech samples were produced by the same talker or by two different individuals, the listeners had to detect subtle differences in the pronunciation of consonants and vowels, as well as in mean F0, intonation patterns, and voice quality. They then had to decide whether these differences merely reflected ordinary deviations in a single talker's speech production, or whether they were caused by a change of talker.[58] As was pointed out previously (§2.2.1.1), the difficulty of this task lies in the fact that some of the phonetic features associated with the speech of one talker always overlap with those of the speech of other talkers. The listeners (particularly when exposed to speech produced in the facewear conditions) therefore had to adopt a rather loose response criterion and accept some amount of difference among the talkers in each same-talker pair as being consistent with a 'same talker' response, and some amount of similarity among the talkers in each different-talker pair as being coherent with a 'different talker' assessment (see Kreiman & Papcun, 1991). In other words, in order to successfully tell apart two talkers with similar voices, and two different speech samples produced by the same talker, the listeners' perceptual systems had to be capable of distinguishing between two sources of variation, namely the variation between the speech of different talkers (inter-talker variability) and the variation within the speech of an individual talker (intra-talker variability).

The auditory-perceptual and acoustic analysis of the speech material offered valuable clues as to which segmental and non-segmental features the listeners may have used to base their decisions on. To name a few examples, auditory analysis accompanied by visual inspection of spectrograms showed that the talkers' consonant productions differed with respect to the nature of the plosive releases of /t/ and /p/. While most

[58] Incidentally, the listeners never compared two identical samples in the same-talker pairs (see §6.2.1.3). We can therefore rule out the possibility that they simply detected auditory changes from sample to sample, which could have been merely technical in nature.

talkers produced single transients for /t/, one talker was found to produce multiple transients. Or, whereas some talkers produced low-energy bursts and only marginally visible transients for /p/, others produced high-energy bursts and clearly visible transients. The intensity of the bursts of /t/ and /p/ relative to the intensity of the vowel also differed across talkers. While some talkers produced high-energy frication at closure release, the burst intensity for other talkers was quite low. Some amount of inter-talker variation in the VOT of /t/ and /p/ was also observed. Additionally, the intensity of the frication noise of /s/ was considerably higher for one of the talkers than for the remaining talkers, and the energy of /f/ was also higher for some than for others. Regarding the suprasegmental features, it was for example observed that the mean F0 of /ɑː/ was higher for some talkers than for others. No differences (based on auditory and visual analysis) were found in terms of F0 contours. Most notably, the speech productions of one of the talkers (talker D) were more distinct than those of the rest of the talkers, e.g. in that he produced ejective-like plosives, a very high-energy /s/, and the highest mean F0 of /ɑː/, that his formants shifted more strongly in the facewear conditions, and that his voice quality was more distinct than the voice qualities of the rest of the talkers. It was hence to be expected that talker D could be more reliably distinguished from the other talkers.

The following sections will explore in more detail the extent to which the listeners' high performance in the talker discrimination task was degraded when their answers were based on facewear speech, and also to what degree the segmental content of the speech impacted on talker discriminability. Finally, it will be discussed whether the acoustic and perceptual changes to the segments caused by facewear are more likely to be talker-specific or talker-independent.

6.3.1 Facewear effects on talker discrimination

Most interestingly in the broader context of the thesis, the current study revealed that the occlusion of the talker's face/mouth during speech production appears to reduce talker discriminability. The participants could still discriminate the (very short)

speech samples with a high degree of accuracy and reliability even under the degraded listening conditions caused by the two face coverings. However, as a result of the disguises the task became more difficult for the perceivers, and correspondingly more error-prone. There was a reduction in speech processing accuracy and processing speed when facewear was involved. In comparison to the near-ceiling performance in the control condition (93% correct), talker discrimination accuracy dropped by approximately 18% in the helmet and 25% in the tape condition. The result in the control condition accords with the outcome of the study by Andics *et al.* (2007), who found a proportion of 87% successful talker discriminations, and with that of Cutler *et al.* (2011), who report a grand mean of approximately 85% accurate talker discriminations.

Why was talker discriminability reduced in facewear speech? The reduction in response accuracy brought about in the helmet and tape trials compared to the control condition must principally be the consequence of certain acoustic modifications to the speech, and perceptual correlates of this. Evidently, these changes to the signal must have been perceptually prominent enough to complicate the listeners' decisions about which speech samples originated from the same talker.

As was repeatedly discussed in previous chapters, the acoustic (and possibly also auditory) changes to speech caused by the talker wearing a face- or mouth-covering mask may originate from sound energy absorption on the part of the facewear material itself, and/or from interference of the facewear with the talker's speech initiation and articulation. The auditory and acoustic analysis of the speech material found evidence in support of both notions.

Firstly, the intensity of plosive bursts (and transients) and the frication noise of both fricatives markedly dropped for all talkers when they were wearing facewear (especially the helmet), which in all likelihood was the consequence of sound energy absorption. This reduction in overall intensity of the consonants is in line with the findings from the acoustic study discussed in Chapter 4.

Secondly, the tight fit of the helmet around the talker's face, and even more radically, the tape adhered to the talker's mouth/cheeks, most probably triggered certain mechanical constraints to the natural motor activity of the talkers' articulators,

especially the lips and jaw. Auditorily, the perturbation of lip movement may have become apparent in consonants which require lip motion in order to be articulated adequately, such as the bilabial stop /p/ and the labiodental fricative /f/. These articulatory facewear effects may be part of the reason that the talker discrimination rates in the facewear conditions were reduced more strongly for /p/ and /f/ than for /t/ and /s/. Furthermore, the limitation of jaw movement may to some extent account for the observed formant shifts. The most prominent formant shifts were observed in the tape condition, where in particular the mean F1 of /ɑː/ was considerably lower than in the control speech.

In addition, some talkers seem to have actively adapted their speaking behaviour to wearing a face covering, for instance by articulating in a more exaggerated way. This was most notable for talker D, who produced highly energetic fricatives and ejective-like stops. Other talkers appear to have compensated for the facewear by raising their level of vocal effort in order to increase the loudness of their speech (especially in the helmet condition, where auditory feedback would have been additionally altered). This may explain the relatively high performance for /f/ in the helmet condition (despite the low intensity of /f/ and the acoustic absorption caused by the helmet). The increase of vocal effort may also account for the finding that most talkers by trend increased the mean F0 of /ɑː/ in the helmet and tape speech.

In sum, the listeners' task of discriminating between two samples of speech was made much more difficult when facewear was involved. It can be hypothesised that the increase in false discriminations in the helmet and tape conditions resulted either from an *increase* of *within*-talker variability – i.e., less similarity of the samples in a same-talker pair – and from a *reduction* of *between*-talker variability – i.e., higher similarity between the samples in a different-talker pair (or both). In the first case, talker-discriminating information encoded in the speech signal produced by the same talker may have been lost by virtue of the facewear, for which reason the detection of the same-talker pair was compromised. In the second case, two samples of speech produced by different talkers may actually have become more similar, which as a result may have hindered the detection of the different-talker pair.

6.3.2 Consonant effects on talker discrimination

In addition to the finding that facewear compromised talker discrimination, the study presents evidence that the segmental content of speech can impact on talker discriminability. The analysis of the response patterns for each of the six consonants individually showed that the listeners' ability to distinguish between unfamiliar talkers substantially varied as a function of the segmental content. As the reader will recall, the nucleus of the test syllables was /ɑː/ throughout the experiment. The variation in talker discrimination performance can therefore be ascribed to changes to the consonantal onset of the syllables.

Specifically, it was found that some consonants led to a higher rate of correct talker discriminations than other consonants.[59] This suggests that some consonants encoded more indexical information that was beneficial for telling apart two talkers, and thus made a bigger contribution to the listeners' success in distinguishing between talkers. These results are in line with those previously reported by Andics *et al.* (2007) and Cutler *et al.* (2011). The novel aspect of the study is that the effect of speech content on talker discrimination is maintained with facewear speech.

The order of consonants that resulted in the lowest proportion of correct talker discriminations to consonants that yielded the highest mean accuracy was, however, not the same in the control and facewear conditions. The ranking of consonants (from high to low) in the control condition was /t/ > /s/ > /m/ > /f/ > /n/ > /p/, in the helmet condition it was /f/ > /t/ > /s/ > /n/ > /p/ > /m/ (manifested by a drop of 14–25% correct compared to the baseline), and in the tape condition it was /n/ > /s/ > /m/ > /t/ > /f/ > /p/ (indicated by a drop of 19–33%). Note in particular that, in accordance with the literature, alveolar consonants carried the largest amount of talker-specific information in the control condition. These rankings of consonants imply that talker discrimination brought about by a particular consonant varied in the helmet and tape speech. This in turn suggests that the perceptual properties of the

[59] It should be remembered, however, that the performance for all consonants in all facewear conditions was significantly above chance level (50%).

consonants were not equally affected by facewear (see significant facewear x consonant interaction).

Why was response accuracy higher for some consonants in facewear speech? The above findings are easier to understand when one recollects that different consonants are characterised by different articulatory and acoustic properties, and are therefore more or less susceptible to specific articulatory and/or acoustic modifications caused by facewear. The auditory/acoustic analysis of the speech data revealed three major trends in this respect.

Firstly, consonants which exhibit an overall high amount of acoustic energy were, as expected, more resilient to facewear effects than low-intensity consonants. The perceptual effects of facewear-induced sound energy absorption may for this reason have been kept within a limit for high-intensity sounds. For example, the frication noise of the low-intensity fricative /f/ was more affected by acoustic absorption than was the high-intensity /s/ (especially in the helmet condition), and hence yielded lower discrimination scores. Similarly, the overall weaker plosive burst of /p/ was more affected than the stronger /t/ burst, and therefore scored lower.

Secondly, consonants which require precise lip movements to be produced were acoustically and perceptually more affected by facewear than those that do not (or only marginally) involve the lips as an active articulator. This assumption was corroborated, for example, by the lower discrimination rates for the bilabial plosive /p/ than for the alveolar plosive /t/, and also by the rates for the labiodental fricative /f/ relative to those for the alveolar fricative /s/, when the consonants had been spoken through facewear (especially the tape).

Thirdly, oral and nasal consonants were in general differently affected in facewear speech. In the tape condition, the listeners on the whole performed better when nasal consonants were presented, whereas in the helmet condition they scored higher when listening to oral consonants. This outcome becomes more understandable when recalling that the talkers' noses were covered in the case of the helmet, but were unconcealed in the tape condition. Hence, it was still possible for the talkers to fairly reliably produce /m/ and /n/ despite the tape across their mouths. The movement of the lips was restricted by the tape, but since lip motion is less required for the

production of the two nasals than for the production of the oral consonants, and since the airflow through the nasal cavity was sustained, the perceptual effects of the tape were less detrimental in case of the nasals than they were for oral consonants.

6.3.3 Inter-talker variation in facewear speech

In addition to the facewear and consonant effects on talker discrimination described in the previous sections, the study revealed a highly complex relationship between the facewear condition under which the tested speech was produced, the type of consonant that was presented to the listeners, and the specific (combination of) talkers in a particular experimental trial. The statistical analysis of the data showed that the strength of the detrimental effect of facewear on talker discriminability did not only vary across the different consonants, but also varied across talkers (see significant facewear x consonant x pair interaction). Despite the complicated patterns that arose, several interesting trends can be inferred from the data.

For one, some talkers were by and large better distinguished from the remaining talkers (higher discriminability) than were other talkers in the test set (lower discriminability). This outcome is unsurprising; the voices of some talkers are simply more distinct than others and are therefore more easily discriminated. The listeners' success in telling apart the same-talker pair from the different-talker pair was also dependent on which speech samples were presented together within a trial. Naturally, it was harder for the listeners to decide which two speech samples came from the same talker when the two presented talkers were very similar-sounding; the decision was easier when the talkers were perceptually more distinct.

The novel and more interesting implication from the results is that the effect of facewear on the perceptual qualities of consonants, and the impact this change in consonant perception has on talker discrimination based on these consonants, appears to be talker-dependent. This means that the consonants produced by different talkers were not equally affected by facewear. Rather, facewear affected the acoustic and

perceptual properties of the consonants produced by some talkers more than those produced by other talkers. That is, some talkers were more vulnerable to facewear effects than others.

The large degree of variation in the speech/voices of different talkers was confirmed by auditory and acoustic analysis of the speech material. Altogether, it was found that the facewear-triggered changes to certain phonetic features of the talkers' speech had the effect that sometimes *within*-talker variation would be reduced (samples from the *same* talker became more similar), and sometimes it would be enhanced (samples from the *same* talker became more distinct). Similarly, *between*-talker variation would sometimes be reduced (samples from *different* talkers became more similar), and sometimes enhanced (samples from *different* talkers became more distinct).

These variable patterns were most prominently illustrated in the formant and F0 data presented in §6.2.3. Furthermore – to recall just a few of the observations from the previous sections – the intensity of the closure release transients of /t/ and /p/ was reduced for some talkers and enhanced for others (especially in the tape condition). The energy of the /p/ burst decreased for some and increased for other talkers, and the intensity of the frication noise of fricatives (especially /s/) also varied substantially across talkers. Moreover, the first three formants of /ɑː/ produced by different talkers were not equally affected in facewear speech, and the mean F0 of /ɑː/ also varied significantly between talkers.

Finally, some of the observations made while working with the facewear data suggest that the magnitude of facewear effects was also dependent on the fit of the facewear. The fit of the helmet was determined by the size of each talker's head (the same helmet was used for all talkers recorded for the AVFC corpus), and the strength of the adherence of the tape to the talkers' lips and cheeks differed between the talkers (due to personal preferences or external factors, such as facial hair). Both the helmet and tape can be considered as an additional constriction outside the vocal tract, which may have been closer or further away from the talker's lips (tight versus loose fit), thus leaving a wider or narrower channel for the air to escape from the vocal tract. This may explain, for example, the varying intensity patterns (especially in the tape condition).

6.3.4 Summary

In conclusion, and returning to the research questions raised at the outset of this chapter, the current study offered relevant new insights into the effects of facewear on unfamiliar talker discrimination. The auditory and acoustic analysis of the speech material furthermore revealed details about the specific nature of articulatory, acoustic and perceptual facewear effects on consonants and vowels, which are valuable to the thesis as a whole. The observations made of four male talkers speaking through a motorcycle helmet and while their mouths were taped closed underline the theoretical considerations drawn in §2.1.2.

The main findings can be summarised as follows:

- phonetically-untrained listeners can successfully discriminate between unfamiliar talkers based on short CV syllables (on average 78.2% correct)
- the segmental content of speech affects talker discrimination
    - some consonants encode more indexical information, i.e., they offer a greater number of talker-specific cues to successful talker discrimination, than other consonants do (see Table 6.2)
- facewear reduces talker discriminability
    - near-ceiling performance in the control condition (92.6% correct), drop by about 18% in the helmet and 25% in the tape condition (see Table 6.2)
    - some consonants are more resilient to articulatory and/or acoustic modifications caused by facewear than other consonants (dependent on overall intensity as well as manner and place of articulation of the sounds)
    - facewear modifies talker-specific properties of consonants and vowels on both the segmental and suprasegmental levels
- facewear changes the properties of speech in a talker-specific manner
    - facewear increases/reduces intra- and inter-talker variability in the signal
    - some talkers are more vulnerable to facewear effects than other talkers (dependent on external factors, e.g. head size, and deliberate/automatic compensation strategies, e.g. rise in vocal effort or hyperarticulation)

% correct talker discrimination							
facewear	**/t/**	**/p/**	**/s/**	**/f/**	**/n/**	**/m/**	**mean**
control	96.0	91.5	92.5	91.8	91.7	92.2	**92.6**
helmet	78.1***	70.3***	76.7***	78.3***	74.3***	67.7***	**74.2*** **
tape	66.2***	59.2***	71.0***	66.2***	73.3***	69.8***	**67.6*** **
mean	**80.1**	**73.7**	**80.1**	**78.7**	**79.8**	**76.6**	**78.2**

Table 6.2. Talker discrimination accuracy for all six consonants as a function of facewear (Experiment 5). '***' denotes a significant difference from the corresponding control condition at $p < .001$.

The discussion of Experiment 5 concludes the empirical part of the thesis. The following and final chapter provides a summary of the core findings from Chapters 1 to 6, and highlights the practical relevance of the results in the context of casework carried out by forensic speech scientists. In closing, several ideas and directions for future research on the topic are proposed.

7 Summary and outlook

7.1 Thesis summary

This thesis has explored the effects of forensically-relevant facial concealment on speech. The term ‘facewear’ was introduced to refer to the various types of face-concealing garments and head coverings that are worn by people in common daily communication situations; for work and leisure, or as an expression of religious, social and cultural affiliation. It also denotes the kind of facewear which is frequently encountered during the commission of crimes or in situations of public disorder. **Chapter 1** listed various examples of the face coverings that are of interest in the present context. The first chapter furthermore informed the reader of the general motivations for the research. These were related to the applicability of the research to casework carried out by forensic speech scientists, and with the ongoing political and social discussions about whether to prohibit the wearing of religious attire and other face and head coverings in public places (e.g. courtrooms and classrooms) and during public gatherings (see ‘anti-mask laws’ and ‘*burqa* debates’). Moreover, a case was made for why it is believed that facewear research will be of relevance in the future. The rationale behind this assertion was that the ubiquity of CCTV cameras in public areas, along with the fact that so many people now carry audio- and video-recording devices (smartphones, mobile phones, etc.), means that the use of facewear by individuals perpetrating crimes or participating in civil disturbances is likely to play a role in future forensic phonetic investigations.

Following the practically-oriented remarks about facewear use in general, **Chapter 2** offered a theoretical background for the experimental work presented in later chapters. The chapter included the presentation of the research directions taken in the thesis, an outline of the field of forensic speech science, and an introduction to previous research on the topic. Furthermore, the second chapter highlighted that even though the primary motivation for the study was of a forensic nature, facewear research can be of interest to researchers in related fields, including experimental and social psychology, sociolinguistics, phonetics, pragmatics and acoustics. The study also contributes to research and usability tests which aim to evaluate and improve speech communication between mask wearers in professional environments,

especially those which exhibit high levels of background noise (e.g. surgical masks or hearing protection devices worn in industrial and medical environments).

The subsequent chapters represented the empirical part of the thesis. First of all, **Chapter 3** introduced the design of an auditory-visual database of facewear speech. The corpus consists of high-quality audio and video recordings (taken from three microphone positions and two camera angles) of ten talkers who are reading aloud phonetically-controlled stimulus sentences in a control condition (absence of facewear), and while wearing one of eight types of (potentially) forensically-relevant face coverings. These were two types of balaclavas, a combination of a hooded sweatshirt and a cotton bandana, a motorcycle crash helmet, a *niqāb* (full-face veil), a full-head rubber mask, a surgical mask, and a piece of adhesive tape across the mouth/cheeks. The database provided the test material for all studies presented in the thesis. The reader is invited to use these data for his/her own research or as reference material in casework (please contact the author for obtaining access to the data).

Chapters 4 to 6 were dedicated to five experiments, which had been designed to empirically test facewear effects on consonants. The goal in the following sections is to summarise the main results of the experiments, and relate them back to the research directions presented in Chapter 2. Fundamentally, this thesis dealt with the question: does facewear influence the way that speech is produced, transmitted, and perceived? It was acknowledged in an earlier chapter (§2.1.3) that numerous approaches could have been applied in order to (begin to) answer this question. For the purpose of the thesis it was decided to centre the analysis on spoken English consonants that were produced while the talker's face is disguised by facewear, and specifically, on the way that the consonants are produced, on their acoustic properties, on how well they can be identified by lay listeners, and on how much idiosyncratic (talker-specific) information they convey.

With these goals in mind, the first two experiments (presented in **Chapter 4**) set out to explore the effects of the aforementioned eight types of face coverings on selected acoustic-phonetic parameters of the voiceless fricatives /s ʃ f θ/ (Experiment 1) and plosives /p t k/ (Experiment 2). Fricatives and plosives were chosen for perceptual and acoustic reasons, and in consideration of their relevance as a consonantal

parameter that is quite commonly analysed by forensic speech experts. The study took into consideration a range of intensity and spectral properties of the frication and burst noise (spectral moments, spectral peak), as well as temporal characteristics of the plosives (closure duration, voice onset time). The comparison of named acoustic measures of facewear speech with those taken from control speech provided valuable insights into the acoustic modifications to speech that can be expected when the talker's face is disguised. On the whole, different face masks were found to alter the acoustic-phonetic properties of consonants differently. However, in spite of the large degree of variation in the data, several interesting trends could be observed.

To begin with, the relatively thin and lightweight textiles of the *niqāb* and surgical mask caused the weakest acoustic effects, especially regarding the intensity of the speech sounds. Some minor changes to the spectral properties of fricatives were observed, but these were only prominent (if at all) for the non-sibilants /f/ and /θ/ (which in any case were more affected by facewear, and exhibited more variation across facewear conditions and samples, than the sibilants /s/ and /ʃ/). However, it must not be inferred from these findings that thinner, lighter, or more porous fabrics will always change the acoustics of speech produced through the fabrics to a lesser extent than thicker, heavier, or more densely-woven ones (see also Llamas *et al.*, 2008). For example, when the consonants had been produced through the scarf and the balaclava (no mouth hole), the intensity of the frication/burst noise sometimes even increased slightly compared to the baseline. It was speculated that this may indicate articulatory compensation behaviour on the part of the talkers, who may have spoken more loudly (i.e., with an increased level of vocal effort) in order to counterbalance the acoustic damping effects triggered by the masks.

All things considered, the strongest impact on speech acoustics was noted in the motorcycle helmet, tape, and rubber mask conditions. These three types of facewear (especially the solid shell of the helmet) absorbed acoustic energy (and thus reduced intensity) much more markedly than any of the other tested face coverings. Moreover, these masks most prominently altered the spectral properties of the consonants (and especially those of the non-sibilants). Most notably, the location of the spectral centre of gravity was reduced by around 1–2kHz, and the spectral distribution was more positively skewed (i.e., towards lower frequencies in the

spectrum). In addition, there tended to be an increase in the voice onset time, and especially in the closure duration of plosives. For example, the closure durations of /p/ and /t/ were up to 60ms longer when the stops had been spoken through the tape, helmet, balaclava (no mouth hole), rubber mask, and surgical mask. Such changes to the temporal composition of the consonants could be indicative of a generally more prolonged articulation when the talker was wearing facewear.

While working with the fricative and plosive recordings, the author observed that facewear occasionally brought about formant-like patterns in the spectrum. These appeared to be the result of attenuation of acoustic energy surrounding the bandwidth of the newly-formed formant(s) rather than enhancement of particular formant frequencies, and may in part account for the typical relocation of centre frequencies. Besides this, the data affirmed the assertion put forward by Llamas *et al.* (2008) that acoustic energy in the speech signal, especially in higher frequency bands, will be suppressed when speech is produced while the talker's face is disguised. In facewear speech (compared to control speech), less sound energy seems to be concentrated above the approximate threshold of 5–6kHz, and relatively more energy is found in lower frequency regions. This sound energy migration (a term borrowed from Stanton *et al.*, 1988) may have been the major source for the typical downward shift of the centre of gravity and for the positive skewing of the spectral distribution in facewear speech. The 'artificial' lowering of centre frequencies caused by facewear acting as a low-pass filter was considered to be conceptually similar to the 'telephone effect' established in forensic phonetics (e.g. Künzel, 2001; Byrne & Foulkes, 2004).

In sum, the results of the acoustic-phonetic examination of facewear speech strongly suggest that face coverings have the potential to considerably change intensity, temporal, and spectral characteristics of fricatives and plosives. The observed 'acoustic facewear effects' are likely to be the consequence of a) acoustic absorption (particularly at frequencies above 5–6kHz), and b) active and passive modifications to the talker's 'natural' speech productions. The former most probably produced the prominent centre of gravity shifts, while the latter may have manifested themselves in increased vocal effort *despite* the transmission loss caused by the mask material (see e.g. the intensity increase for the frication and burst noise) or the prolonged articulation (see e.g. the longer plosive closure durations and voice onset times).

The subsequent experiments carried out within the context of this thesis focused on the effects of facewear on speech perception. This aimed at testing whether the facewear-induced acoustic changes to speech negatively impact on speech processing by lay listeners. The first perception study (presented in **Chapter 5**) consisted of two auditory-visual consonant identification experiments. In both experiments, participants were asked to identify syllable-onset consonants (/p b t d k g f v s z ʃ ʒ θ ð m n/) embedded in nonsense CVC syllables spoken phrase-finally in a standardised carrier sentence (*He said [stimulus]*.). The accuracy of consonant identification was compared when stimuli in the form of sound/video recordings were presented in auditory-only (AO) and auditory-visual (AV) formats. Additionally, it was examined whether consonant identification was affected when the speech had been produced while the talker was wearing one of the eight face masks listed earlier. In the first, baseline-establishing experiment (Experiment 3) both the video and audio quality were kept optimal (studio quality). In the follow-up experiment (Experiment 4) the speech stimuli were presented in background noise (8-talker babble) at low signal-to-noise ratios. A signal detection analysis (*d-prime*) and a sequential information analysis (SINFA) were employed to analyse the resultant recognition errors. In this thesis, only the *d-prime* results were presented (owing to lack of space).

The findings from this study suggest that the perception of syllable-onset consonants quite substantially changes when the consonants are produced while the talkers' mouth/nose or whole face is covered by facewear. In Experiment 3 (quiet listening condition) around 8% recognition errors occurred (43 participants x 576 stimuli each = 24,768 consonant responses in total). That is, participants on average identified 92.2% of the consonants correctly. The hit rates ranged from 94.4% in the most favourable experimental condition (control/AV) to 82% in the least favourable condition (tape/AO). In Experiment 4 (speech-in-noise condition) approximately 60% errors were recorded (39 participants x 576 stimuli each = 22,464 consonant responses in total). This shows that consonant recognition accuracy markedly decreased when the consonants were presented in background noise. In this case, the percentage correct scores ranged from 69% (control/AV) to 12.4% (tape/AO).

A closer inspection of the data revealed that the participants were generally better at identifying consonants when they had had additional access to visual speech cues

from the talker's articulating face. The 'AV effect' was negligible in the quiet listening conditions (except when the tape was tested). This indicates that the presentation of the talker's face did not further enhance consonant identification in such cases (performance was already high in the AO conditions). However, as the listening conditions deteriorated with the addition of babble noise to the original soundtracks, the availability of facial information became, as expected, more important to the perceivers. Most interestingly, the AV effect was maintained with facewear speech. This suggests that visual speech information can be extracted even from a partially-disguised face. A more in-depth analysis showed that the AO and AV hit rates varied substantially as a function of facewear type. In fact, the nine facewear conditions (including the control) clustered into three 'classes'. These differed with respect to the occurrence and strength of the AV effect. This in turn could be related to the nature of visual speech information that could still be recovered from the face.

The first class of facewear included the control condition, the balaclava with the mouth hole, and the tape across the talker's mouth. Here, the AV effect was strongest (i.e., the intelligibility gain was highest when the talker's face was presented). It was argued that this was for the most part the result of the talker's mouth region still being visible to the observers. This allowed the participants to extract lip (and in part tongue) movements, which supported in particular the detection of the consonantal place of articulation.

The second class of facewear consisted of the surgical mask, the balaclava without the mouth hole, and the hoodie/scarf combination. Here, a significant AV effect still emerged, but the effect was diminished overall. The weakening of the AV effect was ascribed to the fact that the talker's entire mouth (+ jaw/larynx) region was covered up, for which reason lip-/tongue-reading was no longer possible. Interestingly, however, consonant recognition still greatly improved when the participants could also see the talker (and not just hear the talker's voice). This indicates that despite the absence of lip movements and fine facial detail in the displays, observers could extract helpful visual speech cues from the talkers' disguised faces. For example, consonant identification was possibly enhanced by the visibility of subtle visual cues, such as the 'inflation' of the surgical mask or slight variations in the positioning of

the scarf. The observers may also have benefited from rather coarse visual speech movements, like the opening and closing gestures of the jaw, which informed them about the time-varying characteristics of the visual speech signal (the three masks were relatively close-fitting). This may have drawn attention to 'critical events' in the speech signal and helped the listeners to detect, for example, relevant syllable onsets among the distracting onsets introduced by the 8-talker babble noise.

The third class of facewear included the *niqāb*, the rubber mask, and the motorcycle helmet. In these cases, no AV effect was observed. That is, consonant intelligibility did not improve when the participants had visual access to the talker's concealed faces. This is understandable on the basis that these three face coverings occluded the entire face (except for a small area around the eyes). For this reason all facial (segmental) information was absent (or at least massively compromised).

To sum up, the first speech perception study presented in this thesis provided evidence that consonant identification (in noise) can be greatly affected when the speech sounds are produced through facewear. The strength of the 'auditory-visual facewear effect' was dependent on the nature of the visual speech cues that were still available to the observers. Identification accuracy was particularly promoted when a face was presented from which the viewers could still recover lip movements. When the mouth region was obscured by a mask, accuracy dropped overall. However, there was still a relative improvement in consonant recognition when the (disguised) face was visible to the participants, compared with when they were only exposed to the talker's voice. Perceivers therefore appear to have made effective use of extraoral facial cues to consonant identity (e.g. from jaw motion).

Finally, the second speech perception study (presented in **Chapter 6**) offered valuable new insights into the effects of facewear on the discrimination between unfamiliar talkers. Experiment 5 investigated lay listeners' ability to distinguish between two unknown talkers when all the listeners had available for comparison were isolated CV syllables. The main goals of the study were, firstly, to examine whether talker discrimination is complicated when listeners' decisions are based on facewear speech, and secondly, whether some consonants possess greater talker-discriminating potential than others. The task of the 24 participants in the study was

to make timed decisions about which pair of speech samples – of two pairs presented in each of 432 experimental trials – were produced by the same talker ('two-interval forced-choice' procedure). The speech material was highly controlled (e.g. for amplitude, interstimulus intervals, and the occurrence of a response bias), and consisted of /Cɑː/ syllables with a systematically varying onset (/t p s f n m/). The syllables were produced by four male talkers in the control (no facewear) condition, while wearing the motorcycle helmet, and with a piece of tape across their mouths.

In total, 78.2% of all talker discriminations were correct. The listeners were able to distinguish between the talkers at significantly better than chance level (50%), even under the degraded listening conditions introduced by the helmet and tape. However, in comparison to their near-ceiling performance in the control condition (93% correct), discrimination accuracy dropped by approximately 18% in the helmet and 25% in the tape condition. The reduced rates of correct responses in the two facewear conditions, along with significant delays in response, indicate that talker discrimination became more difficult for the perceiver – and correspondingly more error-prone – when facewear had changed certain articulatory and acoustic properties of the talker's speech.

Furthermore, some consonants led to a significantly higher proportion of correct talker discriminations than other consonants, which suggests that some consonants provided more talker-specific information that was beneficial for keeping apart two talkers than did others. The 'ranking' of consonants was, however, not the same in the control and facewear conditions: control = /t/ > /s/ > /m/ > /f/ > /n/ > /p/, with response accuracy ranging from 96–91.4%; helmet = /f/ > /t/ > /s/ > /n/ > /p/ > /m/, manifested by a drop of 14–25%; and tape = /n/ > /s/ > /m/ > /t/ > /f/ > /p/, indicated by a drop of 19–33%. This indicates that facewear affected the perceptual properties of the consonants, and hence talker discrimination based on these consonants, to different degrees.

The above findings seem plausible when it is borne in mind that a) different consonants are characterised by different articulatory and acoustic features, and are therefore more or less susceptible to articulatory and acoustic facewear effects, and b) different types of facewear affect the acoustic-phonetic properties of speech

differently (as e.g. shown in Chapter 4). Indeed, the auditory/acoustic analysis of the speech material showed that consonants which are characterised by overall high acoustic energy were generally more resilient to facewear effects than low-intensity consonants (e.g., /f/ was more affected than /s/, and /p/ was more affected than /t/). Note that the reduction of the intensity of fricatives and plosive bursts in the helmet and tape conditions reflects the results of the acoustic study presented in Chapter 4.

In addition, facewear effects on consonants were dependent on the particular consonantal manner and place of articulation involved. For example, consonants which require precise lip movements to be produced were by trend more affected acoustically and perceptually than consonants which do not (or less) involve the lips as an active articulator (see e.g. the lower talker discrimination rates for /p/ and /f/ than for /t/ and /s/). The perturbation of 'normal' lip motion can also explain some of the auditory impressions of the speech data, such as that /m/ spoken through the tape often sounded like a labiodental approximant [ʋ] (the tape prevented the lips from forming a complete bilabial closure). Note that the latter mirrors the common misperception of /m/ as /v/ in the AV consonant identification study discussed in Chapter 5. The obstruction of the nose by a mask often resulted in denasality (acoustic absorption of nasal formants). Lastly, oral and nasal consonants were differently affected in facewear speech. In the tape condition (where the nasal airflow is maintained, but lip motion is perturbed), the listeners performed better overall when nasal consonants were presented, whereas in the helmet condition (where the nasal airflow is disrupted, but lip motion is less perturbed) they scored higher when listening to oral consonants.

In line with the observations of the acoustic study, some of the talkers from Experiment 5 seem to have actively adapted their speaking behaviour to wearing a mask by articulating in an exaggerated way (see e.g. the highly energetic fricatives or ejective-like stops produced by one talker). Others appear to have compensated for the face coverings by raising their vocal effort in order to increase the loudness of their speech (as e.g. suggested by the overall higher mean F0). Moreover, it is conceivable that the limitation of jaw movement accounts for the observed formant shifts (largest reductions of mean F1 of /ɑː/ were found in the tape condition). Lastly,

the high degree of variability in the speech data may in part be ascribed to the fit of the masks for individual talkers.

In sum, the second speech perception study presented in this thesis showed that talker discriminability can be greatly compromised when it is based on facewear speech. The study furthermore revealed that some consonants lead to higher talker discrimination rates than others, and therefore seem to possess greater talker-discriminating potential than others. Moreover, some of the facewear-induced changes to the perceptual properties of the consonants appeared to manifest themselves in a talker-specific manner. This means that the acoustic and perceptual properties of speech produced by some talkers were more affected than the corresponding properties of speech produced by other talkers. Put differently, some talkers seem to be more resistant to 'facewear effects' than others. Consequently, facewear appears to have the capacity to both increase and reduce the variability in speech produced by the same talker (within-talker variability) and by different talkers (between-talker variability).

Based on the empirical results of the thesis, we can conclude that facewear has the potential to significantly affect speech production, acoustics and perception. This finding has interesting implications for criminal investigations in which speech produced through facewear is of particular importance. The next section therefore discusses the practical implications of the observed facewear effects on speech in the context of forensic phonetic casework.

7.2 Practical implications

In the forensic speech science literature, the 'masked robbery' is a frequently-cited example of a scenario in which the victim of or witness to a crime could *hear* but not *see* the offender. However, as was noted repeatedly at the outset of the thesis, the potential effects of a face mask on the (ear)witness's perception of the perpetrator's voice/speech, and/or on the acoustic speech signal, have rarely been studied before (with the exception of Llamas *et al.*, 2008; Zhang & Tan, 2008; and Heath & Moore, 2011). This is rather surprising given the direct forensic relevance of the topic, and the relative frequency of forensic cases which involve speech produced under facial disguise. It appears that until now facewear has merely – and one could say prematurely – been considered as incidental information to a case (provided e.g. by the police or an instructing solicitor).

One major objective of this thesis was therefore to demonstrate that face coverings should be treated as more than just background information to a case, and that they act to do more than just conceal the visual appearance of a person. Of course, in a 'typical' forensic context facewear primarily serves as a (deliberate) *visual disguise* of the identity of a person who does not want to be recognised, e.g. from CCTV footage. The implications for *eye*witness testimony are self-evident in this case. From a forensic phonetic point of view, however, it seems justified to go as far as classifying facewear as a form of (presumably non-deliberate) *voice disguise*.[60] The reasoning behind this decision is that facewear will modify the acoustic and perceptual properties of the speech signal. The present research serves to inform forensic practitioners about the specific nature of (some of) the acoustic and perceptual changes to speech that can – and in practice should – be expected when

[60] Other forensically-relevant non-deliberate forms of voice disguise are encountered in situations where 'external' circumstances change the speaker's usual voice and speech patterns. Examples include the speaker's health, adverse recording and channel characteristics, and objects in the mouth (e.g. a cigarette) or in front of the mouth (e.g. a hand or scarf). These examples contrast with deliberate attempts at voice disguise, whereby the speaker consciously tries to falsify or conceal his/her identity, such as by putting on a regional or foreign accent, by modifying pitch, speaking tempo, or voice quality, by pinching the nose, or by holding a bite-block object (e.g. a pencil) in the mouth (see e.g. Künzel, 2000; Hollien, 2002; Clark & Foulkes, 2007; Zhang & Tan, 2008; Hove & Dellwo, 2012).

working on cases that comprise the analysis of speech produced by a speaker whose mouth/nose or entire face was obstructed when the speech material was recorded.

To begin with, forensic speech scientists carrying out casework should be prepared to take a multitude of factors into account when interpreting the results of their *acoustic-phonetic* analysis of case material that involves facial disguise of one form or another.[61] The findings from this thesis imply that the wearing of facewear should indeed be considered as both a *speaker* and *channel* factor.

To recall from Chapter 2, speaker factors are those parameters which bring about differences between speech samples produced by the same speaker and by different speakers (e.g. language, accent, speaking style, distress, health, drug consumption, or voice disguise). The experiments discussed in previous chapters have demonstrated that facewear has the potential to alter the speaker's articulatory behaviour both actively (e.g. through raised vocal effort or hyperarticulation) and passively (e.g. due to perturbations of lip or jaw movement). Modifications to speech production of this kind may subsequently affect the acoustic properties of the produced speech. In the present context, this articulatory-to-acoustic mapping may account for some of the changes to the intensity, temporal, and spectral characteristics of fricatives and plosives (for a summary of results see §4.3.4).

Channel factors, on the other hand, specify the qualitative differences between two speech samples in terms of their technical properties, or of the environmental conditions in which a recording was made or a voice was witnessed. The results of the present research suggest that the mask materials act as an acoustic filter which impedes the transmission of the speech signal and attenuates certain frequency components. For example, acoustic absorption was greatest in frequency regions above 5–6kHz, which in turn led to appreciable centre of gravity shifts in the speech spectra (especially of low-energy sounds, such as non-sibilant fricatives).

[61] The evidential material may arise in form of audio recordings (e.g. from intercepted phone calls, or police interviews) and video footage (e.g. from CCTV surveillance cameras, or personal recording devices).

These experimental findings highlight that forensic practitioners are strongly advised to take the (possible) articulatory and acoustic facewear effects on speech into consideration when they compare the acoustic properties of two speech samples. This task often arises as part of a speaker comparison exercise, or when intending to corroborate auditory judgements of spoken utterances (discussed further below). It is argued here that experts need to understand that the reliability of their measurements can only be enhanced if, in addition to all other known influencing factors, the effects of facial disguise on speech are taken into account.

In this context, it seems beneficial to point the reader towards some of the observations made while the acoustic study presented in Chapter 4 was being conducted. Firstly, pilot experimentation on the plosive data using various filter and pre-emphasis settings revealed exceedingly large differences between the outcome of certain spectral measures (centre of gravity/standard deviation) on speech that was or was not filtered prior to the analysis. This finding urges caution when different settings are used during the recording of an unknown speaker sample and of a suspect sample, or when extracting acoustic features from the samples. It also calls for a detailed account of relevant settings in publications and reports to be routinely provided. Discrepancies between acoustic measures may be falsely attributed to differences between speakers, when in effect they are merely technical in nature.

A second source of variation in acoustic measurements can emerge from the placement of segment boundaries. This holds especially for acoustically complex sounds, such as plosives. As was reported in §4.2.2.3, researchers adopt different criteria when they segment the speech signal into smaller analytical units. Analysts should bear in mind that even slight differences between the timestamps chosen for a particular segment boundary can potentially change the measurement result (e.g. inconsistent VOT owing to varying criteria for marking the voicing onset in the subsequent voiced segment). In the present data, segment boundaries were more difficult to determine when the speech had been produced through a face covering. For example, a common problem encountered when segmenting the plosives was the lack of a distinct transient or burst. On occasion, it was ambiguous whether a spike visible in the spectrogram and/or waveform was actually produced by the speaker, or

whether it was the product of the mask material (especially of the tape and surgical mask) creating additional ‘crackling’ or ‘rattling’ sounds.

Moving on from outlining the relevance of the research to the acoustic examination of speech in forensic phonetic casework, the findings from the speech perception studies are also of great potential to forensic practitioners. A considerable proportion of the practical work carried out by forensic speech scientists consists of the inspection of the (supra)segmental properties of speech through thorough *aural-perceptual* analysis. The factors that are known or expected to influence an expert witness’s performance in this task have previously been termed *listener* factors in the literature. The current research findings once again clearly suggest that the wearing of facewear should be added to the list of known listener factors (see §2.2.2.3).

On the one hand, the experiments presented in Chapter 5 showed that speech intelligibility (especially in the presence of ambient noise) may be interfered with when the speech is produced through a face covering, and specifically, that the identification of syllable-onset consonants is significantly impaired in facewear speech (for a summary of results see §5.6.2). Participants consistently misperceived certain consonants (particularly fricatives) as other consonants; the magnitude of the changes to consonant perception was dependent on the facewear type tested.

These findings have important practical implications for the aural analysis of evidential speech material by forensic experts, for example when they auditorily evaluate the speaker’s pronunciation of consonants for the purpose of speaker comparison or speaker profiling.[62] Moreover, analysts should be aware of the fact that the quality of their impressionistic transcriptions (in the form of orthographic strings or phonetic symbols) may be compromised further when they transcribe utterances that were produced and recorded while the speaker’s face was disguised. As explained in Chapter 2, experts are asked to deliver comprehensive transcriptions

[62] The international survey of forensic speaker comparison practices by Gold & French (2011) showed that all 36 respondents from 13 countries analyse consonants in the course of their forensic phonetic examinations. The authors state that 88% of the experts evaluate the auditory qualities of consonants during casework (p. 300). Furthermore, 82% reportedly examine aspects of timing, and 48% measure the frequencies of energy loci (p. 300).

in cases where the content of an utterance is of particular evidential value, but is ambiguous or difficult to extract even by trained experts.

One major shortcoming of forensic transcripts is that even experienced professionals may have been biased towards hearing certain (possibly more plausible) utterances over others while transcribing the speech (see e.g. Fraser, 2003; Fraser *et al.*, 2011; Fraser & Stevenson, 2014). Analysts need to understand that this bias can lead to substantial transcription errors. Such errors can become pivotal, especially in disputed utterance/questioned content cases. Here, experts are consulted to help to resolve the dispute between two (or more) parties as to what exactly was said in a particular section of a recording. This is commonly done by way of a comparative aural (and potentially also acoustic) analysis. As pointed out in §2.2.1.3, the meaning of an utterance can drastically change when highly contentious words are wrongly transcribed. What is more, even individual consonants (or vowels) as a constituent of a minimal pair (see e.g. French, 1990) can notably modify the speech content. This consideration further illustrates the relevance of the experimental results on the consonant level presented in Chapter 5.

In addition, the study discussed in Chapter 6 revealed that unfamiliar speaker discrimination can become harder for the perceivers when it is based on speech produced through facewear, and that the segmental content of speech (here, consonants embedded in CV syllables) affects speaker discriminability (see §6.3.4). Overall, the listeners' performance in this experiment was high across experimental conditions. Having said this, it should be stressed that this finding must by no means be misinterpreted or generalised. A high performance in speaker *discrimination* does not imply an equally high performance in speaker *identification*. The identification of a person by his/her voice alone is highly prone to error, even when longer speech samples are available and when the listener is familiar with the speaker. Previous research even suggests that the identification of familiar speakers and listeners' discrimination between unfamiliar speakers involve independent cognitive processes. It was found, for example, that speaker identification can be successful when the ability to discriminate between speakers is absent (e.g. due to brain lesions; see Lancker & Kreiman, 1987; or Kreiman & Papcun, 1991). Nevertheless, the findings from the speaker discrimination study are relevant to cases where the expert is

confronted with a (single) speech recording that (potentially) contains two or more speakers. In this scenario, the analyst may be asked to separate out which sections in the recording were spoken by which person (speaker attribution). While a sizeable body of forensic phonetic research has been concerned with speaker recognition (identification) by expert and lay listeners, less is still known about listeners' ability to auditorily discriminate between speakers. The present study gives some valuable pointers in this direction. Future research addressing related questions is encouraged in order to gain further insights into the (likely) performance of expert/lay listeners' in speaker attribution cases, and to ascertain which factors impact upon speaker discriminability.

Finally, forensic speech scientists need to be conscious of the fact that the quality of lay earwitness testimony may be compromised even further when the perceived speech was produced while the speaker's face was concealed by a face covering. The reliability of earwitness statements is at best already questionable, even under the most favourable listening conditions. Facewear can cast further doubt on the reliability of such statements, in which the witness may report being certain about the words that were used, and/or claim that the speaker's voice was that of a familiar person. Until facewear effects are better understood we cannot with any confidence say whether listeners' reports of this kind can be regarded as of equivalent evidential value to those relating to scenarios in which the speaker's face was not disguised.

To sum up, professional forensic speech analysts are advised to treat the wearing of facewear as (yet) another parameter that can affect speech, and more explicitly, as a parameter that has the potential to increase the variability in speech produced by the same speaker and by different speakers. The main objective of this thesis was to provide quantitative experimental data on which forensic speech experts can ground estimates of the influence that facewear may have on the reliability of evidence produced in connection with relevant cases. Despite its practical limitations, it is anticipated that the study can shed some light on the various effects that are likely to occur – on the parts of both the speaker and the listener – when the speech under investigation was produced through facewear. Owing to the high practical relevance of the subject matter, future research in this area is strongly recommended. Some opportunities for future work are suggested in the next section.

7.3 Opportunities for future research

To begin with, the main point of criticism of the current study (from an applied perspective) is likely to be the high degree of control over the experimental procedures and speech materials employed in the experiments, especially in the speech perception studies. The reader may question the ecological validity of the study (see §2.1.3), and challenge the applicability of the present research to real-life forensic settings. Whilst this may be justified to a certain degree, it seems worth emphasising once more that by narrowing down the analysis to the phonemic level, it was possible to extract some of the articulatory, acoustic and perceptual effects on speech caused *by facewear* – and not by other contingent factors. The 'real-life' aspect of the present work *was* that the speech material was elicited from talkers whose mouth or entire face were actually concealed while talking. To strengthen our understanding of facewear effects on speech, future research should include additional factors in the experiments, and ultimately simulate forensically-realistic communication scenarios (e.g. in the form of mock earwitness situations). Several directions for future research are suggested in the following sections.

In terms of acoustic facewear effects, it would be very beneficial to widen our current knowledge of the impact of facewear on speech acoustics by examining additional speech material. The methodology used in the acoustic analysis of voiceless fricatives and plosives, namely the comparison of the acoustic-phonetic properties of facewear speech with those of control (no facewear) speech, proved helpful in this context. Future research could extend the analysis to other consonants (e.g. nasals and voiced sounds), syllable positions, and phonetic environments, as well as to the analysis of vowels. Regarding the latter, preliminary results from the present study suggest that the frequencies of the first three vowel formants might be lowered when the talker is wearing facewear which is covering the mouth and obstructing jaw motion. In addition, the influence of facewear on suprasegmental features of speech, such as voice quality and fundamental frequency, should be studied further in the future. So far we have seen indications of denasality when the talker's nose is covered, and a tendency for F0 to increase in facewear speech. Lastly, as was already pointed out by Llamas *et al.* (2008), it would be worthwhile to

examine in more detail the transmission loss characteristics of different fabrics and materials (in the style of the experiments discussed in §2.3.1). At present it is difficult to discern whether the observed acoustic modifications to speech were the result of changes to the talker's speech productions, or whether they were the consequence of the facewear material acting as an acoustic filter (or indeed both).

It was explained earlier that in order to fully understand how facewear affects the speech communication process, we need to assess its effects on both the talker and the listener. In §2.1.1 the reader was introduced to several viable research directions in this respect. One of them concerned the listener's perception of his/her own voice and of an interlocutor's speech when the listener is wearing 'earwear' (e.g. helmets or hats that cover the ears, noise-cancelling earplugs, hearing protectors, or audio playback devices). In this context, a range of forensically-relevant studies of the effects of ear-concealing facewear on speech and speaker recognition, and on hearing ability more generally, can be foreseen (related research was introduced in §2.3.3). For example, the vastly increased use of smartphones, MP3 players, and other portable audio playback devices in recent years poses potential safety-related problems, such as road accidents caused by pedestrians or cyclists failing to hear traffic noise because they are wearing headphones or hands-free telephone headsets. Scenarios of this kind could potentially lead to court cases where the prosecution might bring a charge of negligence against the person whose ears were covered during the incident. Relatedly, Llamas *et al.* (2008) remark that situations can easily be envisioned where doubt is cast on the reliability of the testimony of a crime witness whose ears were covered while hearing the offender's voice. Incidental information of this kind (ears concealed) might be used, e.g., by the defence in a judicial trial as a way of trying to refute the witness's assertion that the overheard person was the defendant.

Furthermore, it was proposed in §2.1.1 to study more closely whether (and if so, how) a talker whose face is *not* necessarily disguised adapts his/her speaking behaviour when addressing an interlocutor whose face *is* disguised (especially by facewear that hinders eye contact and the extraction of facial expressions). It was noted previously that such adaptations are possibly triggered by certain expectations and biases, or by emotional and attitudinal reactions, towards the person wearing a

particular face covering. The study by Coniam (2005) gave several interesting pointers towards the kind of modifications that might occur, such as reduced speaking volume and rate, clearer articulation, or enhanced use of body language and eye contact (see §2.3.3). Research into similar issues would be of relevance in respect of the ongoing debates about legal bans of the *burqa* from courtrooms and classrooms (see §1.1.2.2), and would also be worthwhile from a sociolinguistic and sociopsychological perspective more generally.

The possibility that certain expectations of and biases towards a mask wearer's speech might indeed exist was affirmed by a brief questionnaire administered as part of the current research. Prior to participating in the AV consonant identification study presented in Chapter 5, the 44 participants in Experiment 3 were presented (only) with pictures of a person wearing the tested facewear (see Figure 7.1) and were asked to evaluate the (anticipated) intelligibility of speech produced by people wearing the particular piece of facewear. They made their choices on a 5-point Likert scale, where '1' corresponded to 'speech is never intelligible' and '5' indicated 'speech is always intelligible'. The responses were subsequently checked for the correct (i.e., not inverted) use of the scale, and were excluded if this was not the case.

The results, averaged across 42 respondents (after two had been excluded), are shown in Figure 7.1. The figure demonstrates that the intelligibility of facewear speech was, with the exception of speech produced through the tape, consistently rated lower before participants took the listening test than after they had completed the test (findings for the first three face coverings shown in the figure corroborate observations made by Llamas *et al.*, 2008). This suggests that the respondents rated facewear speech as less intelligible before they had actually listened to the stimuli (i.e., their answers were purely based on supposition) than after exposure to the stimuli (i.e., after they had experienced facewear speech).

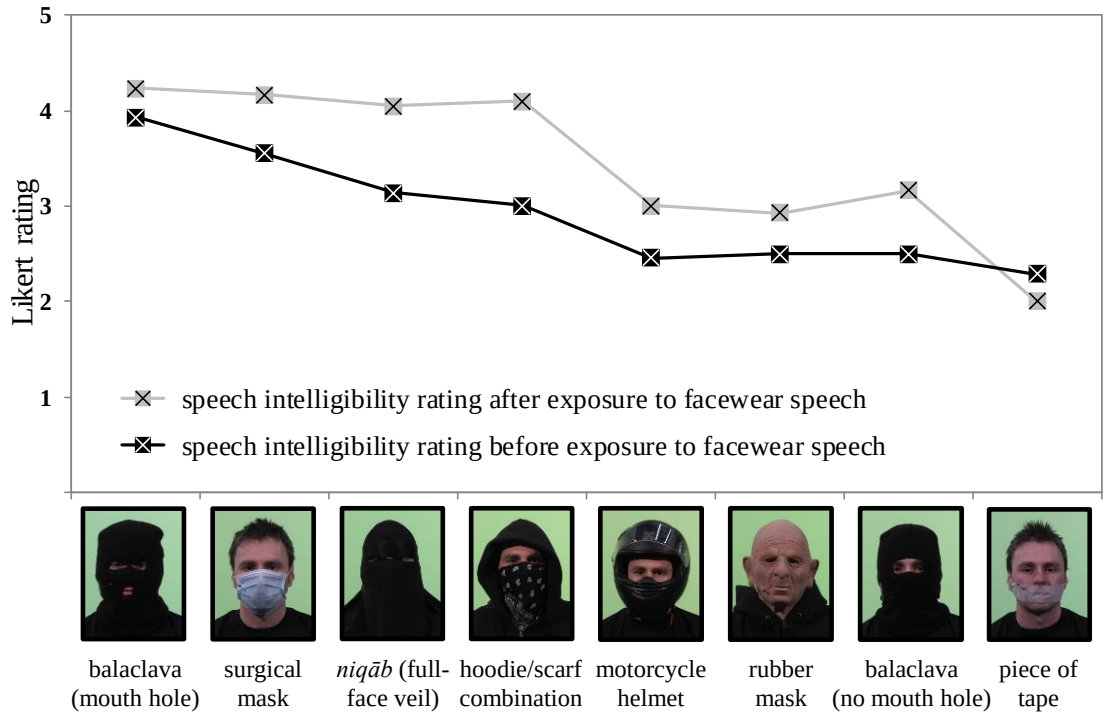

Figure 7.1. Speech intelligibility ratings averaged across 42 respondents. Responses were elicited by means of a 5-point Likert scale (1 = speech is never intelligible, 5 = speech is always intelligible) before and after taking a listening test. Participants rated facewear speech to be less intelligible *before* exposure to facewear speech than after having listened to samples of facewear speech (except tape).

The results from the questionnaire are generally in accordance with preliminary findings reported by Giles (2013), who again isolated the mere 'visual' effects of facewear on the observer. The author compared participants' ratings of speech intelligibility, speech clarity, and perceived 'intimidation' (or threat), when sentences (with or without a 'typically' forensic connotation) were produced while the talker was wearing a balaclava (no mouth hole), a hoodie/scarf combination, a motorcycle helmet, and a *niqāb*. Note that the face coverings used by Giles (2013) were the same items as those tested in the experiments discussed in this thesis. The stimuli were presented in auditory-only, visual-only, and auditory-visual (congruent + incongruent) formats. In the incongruent auditory-visual condition the soundtracks from the control (no facewear) condition were dubbed onto the corresponding videos that showed the talkers' disguised faces. Interestingly, the produced speech was more likely to be perceived as less intelligible, less clear, and more intimidating when the talker was wearing facewear (even when the soundtrack was the same in the AO and AV conditions). These findings highlight once more that research on people's biases

towards (the speech of) mask wearers, and the possible effect on how they perceive and produce speech in response, is advisable.

On a related matter, it would be of great forensic value to study more explicitly observers' emotional reactions towards face coverings that are typically worn in forensically-relevant (and thus potentially highly stressful) situations, like armed attacks or kidnappings. Naturally, most pertinent in this respect are facial guises which may be especially disturbing for the witness to a crime, such as balaclavas. Emotional reactions towards (the wearer of) such masks during exposure to the mask wearer's voice might influence the witness's ability to recognise the voice afterwards and/or to recall from memory the words that were spoken.

The latter assumption is partly based on research by Pickel *et al.* (2003), who report that the 'weapon focus effect' can occur cross-modally. This effect originally referred to the finding that the visibility of a weapon (e.g. a gun or knife) to an eyewitness of a crime can consume attentional resources on the part of the witness, who will allocate attention towards processing the image of the weapon and less towards processing other visible details of the scene (see also Loftus *et al.*, 1987). It has been shown that this impairs the witness's ability to later remember e.g. the offenders' visual appearance. Pickel *et al.* examined whether the presence of a weapon also impairs the witness's memory for *auditory* information. They presented the 'witnesses' in the experiment with a video of an interaction between two characters where the male character was holding a weapon or a neutral object in his hand while talking to the female character. Participants were then asked to recall the content of the male's spoken utterances and to identify him in a mock voice line-up. The authors found that the presence of a weapon impaired the memory for the (semantic) content of the speech, but not for the male's vocal characteristics (e.g. pitch, loudness), and it did not hinder the ability to identify the male's voice in the line-up. On the basis of these results it would be very interesting and informative to adopt a similar experimental design to test whether the presence of a potentially threatening or frightening visual stimulus other than a weapon – a balaclava, say, or another 'typically' forensic facial disguise – harms speech and speaker recognition, in a similar fashion.

In fact, a first attempt in this direction was made by Heath & Moore (2011). As the reader will recall from §2.3.2, they had revealed that a balaclava-concealed face can give rise to the 'face overshadowing effect'. This effect describes the phenomenon whereby the presentation of a picture or video of a face together with the voice at encoding can negatively interfere with the memory and subsequent recognition of the voice (see also Cook & Wilding, 1997, 2001). Contrary to predictions, Heath and Moore found that facial disguise by means of a balaclava does not seem to increase the listener's focus on the voice, and thus does not reduce interference with the visual stimulus. Rather, the presentation of a disguised face limited participants' success in correctly recognising the talker by his/her voice in the same way that an undisguised face did.

The authors suggest that a partly covered-up face still reveals sufficient featural details (e.g. the spatial configuration of the eyes and mouth, or the overall shape of the face) to allow the observer to produce an attentional bias towards the processing of visual over auditory information. Heath and Moore refrain, however, from asserting a possible implication of the results to the effect that *any* visual stimulus can impair the memory for a voice. Indeed, in a previous (unpublished) study by the same authors, a blank disk covering the talker's face did *not* cause the face overshadowing effect. Moreover, Heath & Moore (2011: 138) hypothesise that the balaclava may have induced a sense of personal threat among the observers, and/or introduced an element of 'unusualness' or 'bizarreness'. This may have reduced attentional resources allocated to the voice – i.e., it may have distracted the perceiver from the voice and prioritised scrutiny of the face – in a way that the blank disk did not. The findings of this study emphasise once more that future research on speech processing under facial occlusion conditions should keep the naturalness of the test material as high as possible (as was argued in §5.1.1).

Aside from stressing the naturalness of the facial obstructions, we can hypothesise that the length of exposure to a mask wearer's face may be an important factor. This assumption is based on work by Cook and Wilding (2001), who found that the face overshadowing effect is weaker once observers have become habituated to the face (longer exposure), and by Sheffert & Olson (2004), who also report that experience of a talker's face determines how well the talker's voice can be memorised. It would

be very insightful to carry out more experiments which include the length of exposure to a (facewear-concealed) face as an independent factor. In authentic forensic situations this can be a few seconds up to several hours or even days.

The concept of attentional dominance in the visual processing channel, which is backed up by the work on the face overshadowing effect, is further supported by studies which showed experimentally that interference between auditory and visual information is asymmetrical during speech processing. Experiments by Stevenage *et al.* (2011, 2012, 2013), for example, suggest that 'voice processing' is significantly interfered with by the presence of a face during encoding, but that face processing is not (or much less) impaired when accompanied by a voice. Equally, McAllister *et al.* (1993) tested participants who could see *and* hear the 'criminal' in a mock crime, and who later had to try to recognise the criminal either in a visual line-up or in a voice line-up. They found that visual information about the talker's face interfered with 'voice identification' to a larger degree than auditory information about the talker's voice interfered with face recognition.

Having said that, the reader should also be advised that the above findings to some extent conflict with the results of studies which have shown that voice processing may *benefit* from the co-presentation of a face (see e.g. Stevenage *et al.*, 2011, for further references). Sheffert & Olson (2004), for example, report that the participants in their study were better at recognising the talker's voice after they had learned to recognise the talker's face from video displays which presented both the talker's voice and the talker's articulating face. Preliminary experimentation by the same authors with partly-concealed faces showed that interference of facial information with 'voice learning' does not appear to be the consequence of *facial identity* learning. Rather, it seems to be the result of learning of the *visible speech gestures* from the talker's articulating face, namely those gestures which also help observers to recover the speech content (see Chapter 5).[63] It would be very interesting to test

[63] Specifically, Sheffert & Olson (2004) concealed the portion of the video displays that showed the talker's mouth, lower cheeks and jaw region (but not the eyes, nose and hair). Participants could still learn to identify the faces, at the same accuracy levels that were obtained for the half-face and full-face displays. Intriguingly, however, the presentation of the partly-occluded faces no longer interfered with voice learning. The findings from this

whether similar results emerge for naturally-disguised faces, i.e., whether the presentation of facewear-concealed faces positively or negatively affects memory for (and subsequent recognition of) a voice.

Even though the findings from previous studies are in part inconsistent, it is certain by now that auditory-visual associations during speech processing are not arbitrary. Auditory and visual information derived from natural speech overlap to some degree, and face and voice processing thus interfere with each other in one way or another. This cross-modal linkage of voices and faces (and hence of a person's identity) demands more attention from the forensic phonetic community, especially in light of the fact that earwitnesses to a crime are so often also eyewitnesses. While it seems evident that we rely, under normal circumstances, on our visual judgement to identify a person, it is less clear at this stage to what extent the processing of speech (both for meaning and for the talker's identity) is affected when we simultaneously see and hear the person. We still do not know whether '[r]esearch that is conducted on earwitnesses in the absence of visual information may [...] generalize to ear-witness situations where visual cues are also available' (McAllister *et al.*, 1993: 169).

By and large, the literature suggests that speech and speaker recognition are more accurate when the talker's voice is made salient and visual information is absent during initial exposure to the voice. However, from a forensic point of view we cannot simply infer from this finding that less attention is paid to the voice when the ability to see, say, the perpetrator of a crime is enhanced (and vice versa). We also do not know whether the above findings from (mainly) psycholinguistic studies apply to 'real-life' communication situations, in particular to those in which the offender's face is concealed by facewear. The combination of ear- and eyewitness research, which ultimately includes the possibility that the speaker was wearing a face covering during the incident of interest, is for these reasons highly recommended.

---

study are generally in line with those from Kamachi *et al.* (2003), who report that observers can match a video of an unfamiliar face to an unfamiliar voice, and vice versa. Here, participants were firstly familiarised with a voice, and then presented with two unfamiliar faces. Their task was to decide which face corresponded to the person whose voice they had heard before. Participants were also tested with initially familiarised faces, and their subsequent ability to match the faces to unknown voices. The results of this experiment showed that participants were capable of matching the identity of an unfamiliar person across modalities at levels far from perfect, but significantly above chance.

7.4 In conclusion

This thesis offered an investigation into the effects of forensically-relevant facial concealment on speech acoustics and perception. The key findings are threefold:

Facewear can change speech production. The modifications to speech production may be involuntary on the part of the talker, such as when facewear disrupts nasal airflow or constraints the motor activity of the talker's active articulators (e.g. the lips or jaw). Mask wearers may also deliberately adjust their speaking behaviour so as to compensate for speech perturbations and the anticipated effects on speech acoustics and perception, for example by increasing their vocal effort or speaking in a more prolonged or even exaggerated manner.

Facewear can change speech acoustics. Firstly, even minor modifications to speech production potentially alter the acoustic-phonetic properties of the produced speech signal. Secondly, given that the propagation of the sound wave is hindered when a mask is concealing a talker's mouth/nose, the acoustic energy of certain spectral components of the signal may be attenuated or filtered out. The transmission loss characteristics of different facewear materials will be dependent upon the sound-absorbing properties of the particular material, and to some degree upon other external factors (such as the fit of the mask around the talker's head/face).

Facewear can change speech perception. The modifications to speech production and acoustics can have considerable perceptual consequences for the listener, who can perceive even fine-grained changes to individual speech sounds. Speech recognition (at the consonant level) and talker discriminability may be compromised when it is based on 'facewear speech'. Furthermore, information about visual speech gestures and facial identity will be impoverished when a talker's face is disguised. The absence of facial speech cues can impair auditory-visual speech processing.

In conclusion, then, this thesis has shown that facewear can influence the way that speech is produced, transmitted and perceived. The observed articulatory, acoustic, and auditory-(visual) facewear effects have important implications for legal cases in which speech produced through a face covering is of central relevance. It is therefore strongly recommended that forensic speech experts take these effects into account

when carrying out casework. Future research on the influence of facial concealment on speech acoustics and perception, which can help to fill existing gaps in our understanding of how auditory and visual information interact during natural speech processing, is also strongly encouraged.

A Appendices

A Excerpts from ‘anti-mask’ legislation

An article published in the Harvard Law Review (2004: 2777) quotes:

> Masks can be a powerful aid to unpopular speech. For those who wish to convey messages that are likely to offend governments or individuals, the anonymity that masks provide may encourage the uninhibited expression of views by offering security against reprisal. The masks themselves may also convey a message to observers. Masks can, however, serve illicit ends: the mask wearer may take advantage of the anonymity by committing serious crimes. The enactment of anti-mask laws, which criminalize the public wearing of masks in various contexts, may thus be a sensible anticrime measure.

In the United States, anti-mask laws have their seeds in the Enforcement Act of 1870, which was originally enacted to prevent criminal activities among Ku Klux Klansmen. Section 6 of the Act criminalises visual disguise with the intent to violate another person’s civil rights:

> And be it further enacted, That [sic] if two or more persons shall band or conspire together, or go in disguise upon the public highway, or upon the premises of another, with intent to violate any provision of this act, or to injure, oppress, threaten, or intimidate any citizen [...] shall be held guilty of felony, and, on conviction thereof, shall be fined or imprisoned, or both, at the discretion of the court [...]. (Enforcement Act of 1870, Section 6)

Today, legislation about facial disguise in public is controlled by the state laws of each U.S. state separately, and hence it varies widely across U.S. jurisdictions (Simoni, 1992). Most interesting in the present context is the following paragraph taken from the North Carolina General Statutes. Section 12.8 of Chapter 14 (Criminal Law) of the statutes implies that facewear conceals the identity of a person not only visually, but also by disguising the person’s voice:

> Wearing of masks, hoods, etc., on public property.
> No person or persons shall in this State, while wearing any mask, hood or device whereby the person, face or voice is disguised so as to conceal the identity of the wearer, enter, or appear upon or within the public property of any municipality or county of the State, or of the State of North Carolina. (North Carolina General Statutes, § 14-12.8)

Similar legislation is in place in many other countries around the globe. Most recently in June 2013, Bill C-309 became Canadian law. It makes illegal the concealment of one's identity by means of face masks during 'unlawful assembly'. Subsection 65(2) of the Criminal Code of Canada now reads as follows:

> Every person who commits an offence under subsection (1) while wearing a mask or other disguise to conceal their identity without lawful excuse is guilty of an indictable offence and liable to imprisonment for a term not exceeding 10 years. (Criminal Code of Canada, Subsection 65(2))

In the United Kingdom, the Crime and Disorder Act 1998 (Chapter 37, Part I, Chapter III, Section 25) states:

> Powers to require removal of masks etc. [...]
> (4A) This section also confers on any constable in uniform power—
> (a) to require any person to remove any item which the constable reasonably believes that person is wearing wholly or mainly for the purpose of concealing his identity [...]. (Crime and Disorder Act 1998, Chapter 37, Part I, Chapter III, Section 25)

In Germany, the 'Vermummungsverbot' forbids individuals from disguising their faces in public or carrying any items which prevent identification. Section 2 of § 17a of the 'Versammlungsgesetz' (law governing the right to assembly) declares:

> Es ist auch verboten,
> 1. an derartigen Veranstaltungen in einer Aufmachung, die geeignet und den Umständen nach darauf gerichtet ist, die Feststellung der Identität zu verhindern, teilzunehmen oder den Weg zu derartigen Veranstaltungen in einer solchen Aufmachung zurückzulegen [...].[64] (Versammlungsgesetz, §17a, Absatz 2)

In Sweden, interestingly, it is forbidden to cover the face during public gatherings, but the law specifically *excludes* facewear worn on grounds of religious faith (see Lag (2005:900)).

[64] Translation to English: 'It is also forbidden, 1. to take part in such events while wearing attire which is suitable, and in that context intended, for concealment of the wearer's identity, or to travel to the place of the event wearing such attire [...].'

B Accompanying ethics documentation

B.1 Information sheet (AVFC corpus)

THE UNIVERSITY of York

Information Sheet

'Multimodal Speech and Speaker Recognition' Research Project

Who is involved?

- Research Team:
  - Principal Researcher: Natalie Fecher, PhD Candidate, Marie Curie Initial Training Network 'BBfor2'
  - Supervisor: Dr. Dominic Watt, Senior Lecturer, Department of Language and Linguistic Science
- Ethical Approval: Humanities and Social Sciences Ethics Committee, University of York (Referee: Prof. Helen Weinstein, Acting Chair, email: misc519@york.ac.uk)

What is the study about?

The purpose of the study is to investigate the effects of different types of forensically-relevant face coverings on speech. Specifically, we are interested in what happens to the acoustic speech signal when the speaker is wearing a range of face-concealing garments and headgear. Moreover, we aim to explore whether the listener's perception of speech produced through a face covering is modified in one way or another. The outcome of this research feeds directly into authentic casework carried out by forensic speech scientists.

What does the study involve?

If you decide to participate in the study you will be seated in front of a PC screen and read aloud a list of nonsense syllables which are always embedded in the same carrier sentence, namely 'He said X.'. All syllables will have the same phonetic structure, and will be presented to you using IPA symbols. Examples will be given to you before the start of the recordings. You will repeat the list nine times, once without wearing a face covering, and eight times while wearing one of the following face coverings: a balaclava with a mouth hole, a balaclava without a mouth hole, a surgical mask, a *niqāb* (full-face veil), a combination of a hoodie and a scarf wrapped around your neck/mouth, a full-face rubber mask, a motorcycle crash helmet, and a strip of tape adhered gently to your mouth/cheeks. Further instructions will be provided to you prior to the recordings. There will also be a short training session during which you can familiarise yourself with the experimental procedure.

What kinds of recordings will be made of me?

We will make audio recordings (with three microphones, placed at various distances from you) and video recordings (with two cameras, one placed in front of you and one placed to your side).

May I take a break?

You may take as many breaks as you like. Please note that the task can be quite demanding as you will have to read the same list of sentences nine times. But we will offer free refreshments to keep you going!

Where will the study take place?

The recordings will take place in a recording studio at the Department of Theatre, Film and Television, University of York, Heslington, York, YO10 5DD.

How much time will the study take?

Participation will take about 2 hours and 30 minutes in total.

Department of Language and Linguistic Science, University of York, Heslington,
York, YO10 5DD, United Kingdom. Tel: +44 (0)1904 432650. Email: natalie.fecher@york.ac.uk.

THE UNIVERSITY of York

Information Sheet

Will I be paid for participating?

Yes, you will be paid £25.00 for your participation.

What happens to the data?

All data will be held by the Department of Language and Linguistic Science in accordance with the 1998 Data Protection Act. All data produced in the study will be kept and transferred anonymously and treated strictly confidentially. Only the above-named researchers will have access to the information. Anonymous data will be kept for a minimum of three years, which is the time period of the aforementioned 'BBfor2' research network (http://bbfor2.net/). After that, all personal information will be destroyed. Data from this study may also be used in conjunction with research by other network members, but only with permission of the principal investigator.

What happens to the results of the study?

A report of the study may be submitted for publication, but individual participants will not be identifiable in such a report. If you wish to receive information about your personal test results, or the outcome of the project as a whole, you can contact us at any time. Contact details are given further below.

Can I withdraw from the study?

Participation is entirely voluntary. You are not obliged to be involved. If you decide to participate you can withdraw at any time without giving any reason and without any consequences. If you are a university (under)graduate student, withdrawal from the research will not prejudice your future academic progress in any way.

Who can I contact for more information?

Please contact the principal researcher, Natalie Fecher, for further information. If you have any queries about the research please do not hesitate to contact her at the contact details given at the bottom of this page.

Thank you for reading this information sheet!

Department of Language and Linguistic Science, University of York, Heslington,
York, YO10 5DD, United Kingdom. Tel: +44 (0)1904 432650. Email: natalie.fecher@york.ac.uk.

B.2 Consent form (AVFC corpus)

THE UNIVERSITY of York

Consent Form

'Multimodal Speech and Speaker Recognition' Research Project

Involved:

- Research Team:
  - Principal Researcher: Natalie Fecher, Ph.D. Candidate, Marie Curie Initial Training Network 'BBfor2'
  - Supervisor: Dr. Dominic Watt, Senior Lecturer, Department of Language and Linguistic Science
- Ethical Approval: Humanities and Social Sciences Ethics Committee, University of York (Referee: Prof. Helen Weinstein, Acting Chair, email: misc519@york.ac.uk)

This form is to state whether or not you agree to take part in the study. Please read and answer the next questions. If there is anything you don't understand, or if you want more information, please ask.

Have you read and understood the information sheet about the study?	Yes ❒	No ❒
Have you had the opportunity to ask questions about the aims and procedures of the study?	Yes ❒	No ❒
Do you understand that the information you provide will be held in confidence by the research team?	Yes ❒	No ❒
Do you agree that the data you provide may be used in future research?	Yes ❒	No ❒
Do you agree that we make audio and video recordings of you?	Yes ❒	No ❒
Do you agree to take part in the study?	Yes ❒	No ❒
Do you understand that you may withdraw from the study at any time and for any reason, without affecting any services you receive?	Yes ❒	No ❒
Do you want to be informed about the results of the study?	Yes ❒	No ❒

All data will be held by the Department of Language and Linguistic Science in accordance with the 1998 Data Protection Act. Data will be kept and transferred anonymously and treated confidentially.

Your name (in capitals): _____

Your signature: _____

Principal researcher's signature: _____

Date: _____

Department of Language and Linguistic Science, University of York, Heslington,
York, YO10 5DD, United Kingdom. Tel: +44 (0)1904 432650. Email: natalie.fecher@york.ac.uk.

B.3 Information sheet (Experiment 3)

THE UNIVERSITY of York

Information Sheet

'Multimodal Speech and Speaker Recognition' Research Project

Who is involved?

- Research Team:
  - Principal Researcher: Natalie Fecher, PhD Candidate, Marie Curie Initial Training Network 'BBfor2'
  - Supervisor: Dr. Dominic Watt, Senior Lecturer, Department of Language and Linguistic Science
- Ethical Approval: Humanities and Social Sciences Ethics Committee, University of York (Referee: Prof. Helen Weinstein, Acting Chair, email: misc519@york.ac.uk)

What is the study about?

The purpose of the study is to investigate auditory-visual speech recognition in a forensic context. When people try to understand what another person is saying they use different types of information available. In this project we explore how this information is combined during speech recognition. This is done by introducing forensically-relevant communication situations, which involve speakers who are wearing a variety of face coverings, such as balaclavas, motorcycle helmets, or surgical masks.

What does the study involve?

If you decide to participate in the study you will carry out a computer task for which you will be seated in front of a PC screen while wearing headphones. You will listen to audio and video recordings of different speakers and report judgements about the words presented to you. More specifically, you will always hear/watch the speakers saying the same sentence, namely 'He said X.', where 'X' represents a series of nonsense syllables containing target consonant sounds. You will make judgements about what consonants you heard in the words by clicking on the appropriate symbols on the screen. Detailed instructions will be provided to you prior to the experiment. There will also be a short training session during which you can familiarise yourself with the procedure.

May I take a break?

You may take as many breaks as you like. Please note that the task can be quite demanding as you will have to listen to many recordings in a row. But we will offer free refreshments to keep you going!

Where will the study take place?

The study will take place in a computer laboratory at the Department of Language and Linguistic Science, University of York, Heslington, York, YO10 5DD.

How much time will the study take?

Participation will take about 1 hour 30 minutes in total.

Will I be paid for participating?

Yes, you will be paid £8 for your participation.

Department of Language and Linguistic Science, University of York, Heslington,
York, YO10 5DD, United Kingdom. Tel: +44 (0)1904 432650. Email: natalie.fecher@york.ac.uk.

THE UNIVERSITY of York

Information Sheet

What happens to the data?

All data will be held by the Department of Language and Linguistic Science in accordance with the 1998 Data Protection Act. All data produced in the study will be kept and transferred anonymously and treated strictly confidentially. Only the above-named researchers will have access to the information. Anonymous data will be kept for a minimum of three years, which is the time period of the aforementioned 'BBfor2' research network (http://bbfor2.net/). After that, all personal information will be destroyed. Data from this study may also be used in conjunction with research by other network members, but only with permission of the principal investigator.

What happens to the results of the study?

A report of the study may be submitted for publication, but individual participants will not be identifiable in such a report. If you wish to receive information about your personal test results, or the outcome of the project as a whole, you can contact us at any time. Contact details are given further below.

Can I withdraw from the study?

Participation is entirely voluntary. You are not obliged to be involved. If you decide to participate you can withdraw at any time without giving any reason and without any consequences. If you are a university (under)graduate student, withdrawal from the research will not prejudice your future academic progress in any way.

Who can I contact for more information?

Please contact the principal researcher, Natalie Fecher, for further information. If you have any queries about the research please do not hesitate to contact her at the contact details given at the bottom of this page.

Thank you for reading this information sheet!

Department of Language and Linguistic Science, University of York, Heslington,
York, YO10 5DD, United Kingdom. Tel: +44 (0)1904 432650. Email: natalie.fecher@york.ac.uk.

B.4 Consent form (Experiment 3)

THE UNIVERSITY of York

Consent Form

'Multimodal Speech and Speaker Recognition' Research Project

Involved:

- Research Team:
  - Principal Researcher: Natalie Fecher, Ph.D. Candidate, Marie Curie Initial Training Network 'BBfor2'
  - Supervisor: Dr. Dominic Watt, Senior Lecturer, Department of Language and Linguistic Science
- Ethical Approval: Humanities and Social. Sciences Ethics Committee, University of York (Referee: Prof. Helen Weinstein, Acting Chair, email: misc519@york.ac.uk)

This form is to state whether or not you agree to take part in the study. Please read and answer the next questions. If there is anything you don't understand, or if you want more information, please ask.

Have you had the opportunity to ask questions about the aims and procedures of the study?	Yes ❒	No ❒
Do you understand that the information you provide will be held in confidence by the research team?	Yes ❒	No ❒
Do you agree that the data you provide may be used in future research?	Yes ❒	No ❒
Do you agree to take part in the study?	Yes ❒	No ❒
Do you understand that you may withdraw from the study at any time and for any reason, without affecting any services you receive?	Yes ❒	No ❒
Do you want to be informed about the results of the study?	Yes ❒	No ❒

All data will be held by the Department of Language and Linguistic Science in accordance with the 1998 Data Protection Act. Data will be kept and transferred anonymously and treated confidentially.

Your name (in capitals): _____

Your signature: _____

Principal researcher's signature: _____

Date: _____

Department of Language and Linguistic Science, University of York, Heslington,
York, YO10 5DD, United Kingdom. Tel: +44 (0)1904 432650. Email: natalie.fecher@york.ac.uk.

B.5 Information sheet (Experiment 4)

Human Research Ethics Committee
Office of Research Services

University of Western Sydney

Participant Information Sheet (General)

Project Title: Auditory-visual consonant recognition in speech produced through facewear

Who is carrying out the study?

The study is an international collaborative study coordinated by Natalie Fecher, Dominic Watt (both Department of Language and Linguistic Science, University of York, United Kingdom), and Chris Davis (The MARCS Institute, University of Western Sydney, Australia).

You are invited to participate in a study conducted by Natalie Fecher, Marie Curie Early Stage Researcher, Department of Language and Linguistic Science, University of York, United Kingdom.

What is the study about?

The purpose of the study is to investigate auditory-visual speech recognition in a forensic context. When people try to understand what another person is saying they use different types of information available. In this project we explore how this information is combined during speech recognition. This is done by introducing forensically-relevant communication situations, which involve speakers who are wearing a variety of face coverings, such as balaclavas, motorcycle helmets, or surgical masks.

What does the study involve?

If you decide to participate in the study you will carry out a computer task for which you will be seated in front of a PC screen while wearing headphones. You will listen to audio and video recordings of different speakers and report judgements about the words presented to you. More specifically, you will always hear/watch the speakers saying the same sentence, namely 'He said X.', where 'X' represents a series of nonsense syllables containing target consonant sounds. You will make judgements about what consonants you heard in the words by clicking on the appropriate symbols on the screen. Detailed instructions will be provided to you prior to the experiment. There will also be a short training session during which you can familiarise yourself with the procedure.

How much time will the study take?

Participation will take about 60-90 minutes in total.

Will the study benefit me?

You have the opportunity to take an active part in a large EU-funded research project. By participating we hope that you will gain interesting new insights into a less known domain in the forensic sciences (compared to DNA analysis or fingerprinting), namely forensic speech science. The outcome of the research carried out in this field of study feeds directly into authentic casework carried out by forensic-phonetic experts. Dependent on your background, participation may be of personal and potentially of

professional benefit to you as you can learn more about this very practically-oriented area of research. If you are a university (under)graduate student you can find out more about the design and set-up of a research experiment, as well as the kind of highly specialised questions asked in academic research.

Will the study involve any discomfort for me?

No. The study does not involve the risk of harm or discomfort in any respect.

How is this study being paid for?

The study is sponsored by the Marie Curie Initial Training Network 'Bayesian Biometrics for Forensics (BBfor2)', which receives funding through the European Commission's Seventh Framework Programme (FP7) under grant agreement number 238803. If you are a first year psychology student you will get course credit for your participation.

Will anyone else know the results? How will the results be disseminated?

All data produced in the study will be kept and transferred anonymously and treated strictly confidentially. Only the aforementioned researchers will have access to the information. A report of the study may be submitted for publication, but individual participants will not be identifiable in such a report. If you wish to receive information about your personal test results, or the outcome of the project as a whole, you can contact us at any time. Contact details are given further below.

Can I withdraw from the study?

Participation is entirely voluntary. You are not obliged to be involved. If you decide to participate you can withdraw at any time without giving any reason and without any consequences. If you are a university (under)graduate student, withdrawal from the research will not prejudice your future academic progress in any way.

Can I tell other people about the study?

Yes, you can tell other people about the study by providing them with the chief investigator's contact details. They can contact the chief investigator to discuss their participation in the research project and obtain an information sheet.

What if I require further information?

When you have read this information, Natalie will discuss it with you further and answer any questions you may have. If you would like to know more at any stage, please feel free to contact Natalie Fecher, Marie Curie Early Stage Researcher (email: natalie.fecher@york.ac.uk, phone: +44(0)1904 432 668).

What if I have a complaint?

This study has been approved by the University of Western Sydney Human Research Ethics Committee. The approval number is H9496.

If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Office of Research Services on Tel +61 2 4736 0229 Fax +61 2 4736 0013 or email humanethics@uws.edu.au.

Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.

If you agree to participate in this study, you may be asked to sign the Participant Consent Form.

B.6 Consent form (Experiment 4)

Human Research Ethics Committee
Office of Research Services

University of Western Sydney

Participant Consent Form

Project Title: Auditory-visual consonant recognition in speech produced through facewear

I,……....……, consent to participate in the research project titled 'Auditory-visual consonant recognition in speech produced through facewear'.

I acknowledge that:

I have read the participant information sheet and have been given the opportunity to discuss the information and my involvement in the project with the researcher/s.

The procedures required for the project and the time involved have been explained to me, and any questions I have about the project have been answered to my satisfaction.

I understand that my involvement is confidential and that the information gained during the study may be published but no information about me will be used in any way that reveals my identity.

I understand that I can withdraw from the study at any time, without affecting my relationship with the researcher/s now or in the future.

Signed: _____

Name: _____

Date: _____

Return Address:

This study has been approved by the University of Western Sydney Human Research Ethics Committee.

The Approval number is: H9496

If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Office of Research Services on Tel +61 2 4736 0229 Fax +61 2 4736 0013 or email humanethics@uws.edu.au. Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.

C AVFC corpus documentation

C.1 Questionnaire

THE UNIVERSITY of York

Session and Speaker Information

General information *(to be completed by the experimenter)*

Session name: _____

Staff: _____

Date of session: _____

Start/end of session: _____

Recording environment: _____

Recording equipment: _____

_____

_____

Comments/difficulties: _____

_____

_____

Demographical data *(to be completed by the participant)*

Surname, forename: _____

Contact details (tel/email): _____

Gender: ❒ female ❒ male

Age/year of birth: _____

Background information

Place of birth: _____

Place of primary school: _____

Place(s) of residence: _____

_____

Education: _____

Native language: _____

Second language (if appl.): _____

Dialect/accent: _____

_____

Do you *regularly* wear any kind of face-concealing garment or head cover for occupational, recreational, or religious reasons? If yes, please specify which ones and how often you wear them:

❒ No, never.

❒ Yes: _____

_____

Department of Language and Linguistic Science, University of York, Heslington,
York, YO10 5DD, United Kingdom. Tel: +44 (0)1904 432650. Email: natalie.fecher@york.ac.uk.

C.2 Reading passage

The boy who cried wolf

There was once a poor shepherd boy who used to watch his flocks in the fields next to a dark forest near the foot of a mountain. One hot afternoon, he thought up a good plan to get some company for himself and also have a little fun. Raising his fist in the air, he ran down to the village shouting 'Wolf, Wolf.' As soon as they heard him, the villagers all rushed from their homes, full of concern for his safety, and two of his cousins even stayed with him for a short while. This gave the boy so much pleasure that a few days later he tried exactly the same trick again, and once more he was successful. However, not long after, a wolf that had just escaped from the zoo was looking for a change from its usual diet of chicken and duck. So, overcoming its fear of being shot, it actually did come out from the forest and began to threaten the sheep. Racing down to the village, the boy of course cried out even louder than before. Unfortunately, as all the villagers were convinced that he was trying to fool them a third time, they told him, 'Go away and don't bother us again.' And so the wolf had a feast. (This text was extracted from Deterding, 2006: 193.)

C.3 Recording protocol

	list	facewear		list	facewear
Male 1	9	control	**Female 1**	1	control
	2	hoodie/scarf combination		9	balaclava (no mouth hole)
	3	rubber mask		2	*niqāb* (full-face veil)
	4	surgical mask		3	hoodie/scarf combination
	5	balaclava (mouth hole)		4	rubber mask
	6	motorcycle crash helmet		5	surgical mask
	7	balaclava (no mouth hole)		6	balaclava (mouth hole)
	8	*niqāb* (full-face veil)		7	motorcycle crash helmet
	1	strip of adhesive tape		8	strip of adhesive tape
Male 2	5	control	**Female 2**	8	control
	6	rubber mask		9	balaclava (no mouth hole)
	7	surgical mask		1	*niqāb* (full-face veil)
	8	balaclava (mouth hole)		2	hoodie/scarf combination
	9	motorcycle crash helmet		3	surgical mask
	1	balaclava (no mouth hole)		4	rubber mask
	2	*niqāb* (full-face veil)		5	balaclava (mouth hole)
	3	hoodie/scarf combination		6	motorcycle crash helmet
	4	strip of adhesive tape		7	strip of adhesive tape
Male 3	2	control	**Female 3**	6	control
	3	motorcycle crash helmet		7	*niqāb* (full-face veil)
	4	surgical mask		8	hoodie/scarf combination
	5	balaclava (no mouth hole)		9	rubber mask
	6	*niqāb* (full-face veil)		1	surgical mask
	7	hoodie/scarf combination		2	balaclava (mouth hole)
	8	rubber mask		3	motorcycle crash helmet
	9	balaclava (mouth hole)		4	balaclava (no mouth hole)
	1	strip of adhesive tape		5	strip of adhesive tape
Male 4	4	control	**Female 4**	1	control
	5	surgical mask		2	surgical mask
	6	rubber mask		3	rubber mask
	7	balaclava (mouth hole)		4	balaclava (mouth hole)
	8	motorcycle crash helmet		5	motorcycle crash helmet
	9	balaclava (no mouth hole)		6	balaclava (no mouth hole)
	1	*niqāb* (full-face veil)		7	*niqāb* (full-face veil)
	2	hoodie/scarf combination		8	hoodie/scarf combination
	3	strip of adhesive tape		9	strip of adhesive tape
Male 5	3	control	**Female 5**	7	control
	4	surgical mask		8	rubber mask
	5	balaclava (mouth hole)		9	balaclava (mouth hole)
	6	motorcycle crash helmet		1	surgical mask
	7	balaclava (no mouth hole)		2	motorcycle crash helmet
	8	*niqāb* (full-face veil)		3	balaclava (no mouth hole)
	9	hoodie/scarf combination		4	*niqāb* (full-face veil)
	1	rubber mask		5	hoodie/scarf combination
	2	strip of adhesive tape		6	strip of adhesive tape

Table C.1. Recording protocol used during the recording sessions for the 'Audio-Visual Face Cover' (AVFC) corpus. The order of the 64 syllables in the stimulus list was randomised nine times (obtaining lists 1–9). Each talker read aloud all nine lists in random order. The order of facewear conditions was also different for each talker.

D Supplementary results

D.1 Confusion matrices

balaclava (mouth hole), quiet, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	85	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	86
d	0	82	3	0	0	0	0	0	0	0	0	0	1	0	0	0	86
ð	0	0	44	0	0	0	0	0	0	0	0	2	34	6	0	0	86
f	0	0	0	82	0	0	0	0	0	0	0	0	3	1	0	0	86
g	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	84	1	1	0	0	0	0	86
ʃ	0	0	0	0	1	0	0	0	0	2	82	0	0	0	0	1	86
t	0	0	0	0	0	1	0	0	1	0	0	83	1	0	0	0	86
θ	0	0	1	7	0	0	0	0	0	0	0	2	76	0	0	0	86
v	0	0	2	0	0	0	0	0	0	0	0	0	0	84	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	85	1	86
ʒ	0	0	1	0	7	0	0	0	0	0	1	0	0	0	2	75	86
total	85	82	51	89	95	87	86	86	87	86	84	88	115	91	87	77	1376

Table D.1. Confusion matrix for the consonants presented auditorily in the balaclava (mouth hole) condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response					**balaclava (no mouth hole), quiet, AO**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	85	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	1	72	7	0	1	0	0	0	0	0	0	0	5	0	0	0	86
ð	0	0	52	0	0	0	0	0	0	0	1	2	27	3	1	0	86
f	0	0	0	83	0	0	0	0	0	0	0	0	3	0	0	0	86
g	0	0	1	0	84	0	0	0	0	0	0	0	0	0	0	1	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	85	0	0	1	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	82	1	0	0	0	3	0	86
ʃ	0	0	0	0	0	0	0	0	0	3	83	0	0	0	0	0	86
t	0	0	0	0	0	0	0	0	0	0	0	82	4	0	0	0	86
θ	0	0	1	10	0	0	0	0	1	0	0	1	73	0	0	0	86
v	0	0	2	1	1	0	0	0	0	0	0	0	0	82	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	82	4	86
ʒ	0	0	0	0	7	0	0	0	0	0	0	0	0	0	1	78	86
total	86	73	63	94	93	86	86	86	86	85	85	86	112	85	87	83	1376

Table D.2. Confusion matrix for the consonants presented auditorily in the balaclava (no mouth hole) condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

control, quiet, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	82	3	0	0	0	0	0	0	0	0	0	1	0	0	0	86
ð	0	0	47	0	0	0	0	0	0	0	0	2	28	8	0	1	86
f	0	0	0	83	0	0	0	0	0	0	0	0	3	0	0	0	86
g	0	0	0	0	85	0	0	0	0	0	0	0	0	0	0	1	86
k	0	0	0	0	0	85	0	0	0	0	0	0	1	0	0	0	86
m	0	0	0	0	0	0	85	0	1	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	81	2	2	0	0	1	0	86
ʃ	0	0	0	0	0	0	0	0	0	1	85	0	0	0	0	0	86
t	0	0	0	0	0	0	0	0	0	0	0	83	3	0	0	0	86
θ	0	0	1	9	0	0	0	0	0	0	0	2	73	1	0	0	86
v	0	0	1	0	1	0	0	0	0	0	0	0	0	84	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	83	3	86
ʒ	0	0	1	0	7	0	0	0	0	0	0	0	0	0	0	78	86
total	86	82	53	92	93	85	85	86	87	82	87	89	109	93	84	83	1376

Table D.3. Confusion matrix for the consonants presented auditorily in the control condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

hoodie/scarf combination, quiet, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	1	81	3	0	0	0	0	0	0	0	0	0	1	0	0	0	86
ð	0	0	49	0	0	0	0	0	0	0	1	0	24	12	0	0	86
f	0	0	0	84	0	0	0	0	0	0	0	0	1	1	0	0	86
g	0	1	0	0	85	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	76	3	0	0	0	7	0	86
ʃ	0	0	0	0	0	0	0	0	0	0	85	0	0	0	0	1	86
t	0	0	0	0	0	0	0	0	0	0	0	85	1	0	0	0	86
θ	0	0	1	14	0	0	0	0	0	0	0	2	68	1	0	0	86
v	0	0	6	0	0	0	0	0	0	0	0	0	0	79	1	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	84	2	86
ʒ	0	1	1	0	6	0	0	0	0	0	0	0	0	0	1	77	86
total	87	83	60	98	91	86	86	86	86	76	89	87	95	93	93	80	1376

Table D.4. Confusion matrix for the consonants presented auditorily in the hoodie/scarf condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response					**motorcycle helmet, quiet, AO**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	83	3	0	0	0	0	0	0	0	0	0	0	0	0	0	86
ð	0	0	44	0	0	0	0	0	0	0	0	0	25	12	4	1	86
f	0	0	0	82	0	0	0	0	0	0	0	0	4	0	0	0	86
g	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	80	0	2	0	0	4	0	86
ʃ	0	0	0	0	0	0	0	0	0	1	84	0	0	0	1	0	86
t	0	0	0	0	0	1	0	0	0	0	0	85	0	0	0	0	86
θ	0	0	1	3	0	0	0	0	0	1	0	2	79	0	0	0	86
v	0	0	1	0	0	0	0	0	0	0	0	0	2	83	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	1	82	3	86
ʒ	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	77	86
total	86	83	49	85	95	87	86	86	86	82	84	89	110	96	91	81	1376

Table D.5. Confusion matrix for the consonants presented auditorily in the motorcycle helmet condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

***niqāb*, quiet, AO**

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	83	3	0	0	0	0	0	0	0	0	0	0	0	0	0	86
ð	0	1	50	0	0	0	0	0	0	0	1	1	32	1	0	0	86
f	0	0	1	75	0	0	0	0	1	0	0	0	9	0	0	0	86
g	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	1	85	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	83	1	0	0	0	2	0	86
ʃ	0	0	0	0	1	0	0	0	0	2	77	0	0	0	0	6	86
t	0	0	0	0	0	0	0	0	0	0	0	84	2	0	0	0	86
θ	0	0	3	15	0	0	0	0	0	2	0	3	63	0	0	0	86
v	0	0	2	0	0	0	0	0	0	0	0	0	0	83	1	0	86
z	0	0	0	0	0	0	0	0	0	1	0	0	0	0	80	5	86
ʒ	0	0	0	0	3	0	0	0	0	0	0	0	0	0	3	80	86
total	86	84	59	90	90	86	87	85	87	88	79	88	106	84	86	91	1376

Table D.6. Confusion matrix for the consonants presented auditorily in the *niqāb* condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response					**rubber mask, quiet, AO**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	1	79	4	0	0	0	0	0	0	0	0	0	2	0	0	0	86
ð	0	0	43	0	0	0	0	0	0	0	0	2	35	4	1	1	86
f	0	0	1	77	0	1	0	0	0	0	0	0	4	3	0	0	86
g	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	85	0	0	0	0	1	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	81	2	0	0	0	3	0	86
ʃ	0	0	0	0	0	0	0	0	0	1	84	0	1	0	0	0	86
t	0	0	0	0	0	0	0	0	0	0	0	85	1	0	0	0	86
θ	0	0	4	10	0	0	0	0	0	0	0	2	69	1	0	0	86
v	0	0	2	0	0	0	0	0	0	0	0	1	1	82	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	83	3	86
ʒ	0	0	0	0	8	0	0	0	0	0	0	0	0	0	2	76	86
total	87	79	54	87	94	86	86	86	86	82	87	90	113	90	89	80	1376

Table D.7. Confusion matrix for the consonants presented auditorily in the rubber mask condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response					**surgical mask, quiet, AO**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	83	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	84	2	0	0	0	0	0	0	0	0	0	0	0	0	0	86
ð	0	0	53	0	0	0	0	0	0	0	0	3	24	5	0	1	86
f	0	0	0	81	0	0	0	0	1	0	0	0	3	1	0	0	86
g	0	0	0	0	84	0	0	0	0	0	0	0	0	1	0	1	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	82	1	0	0	0	3	0	86
ʃ	0	0	0	0	0	0	0	0	0	1	85	0	0	0	0	0	86
t	0	0	0	0	0	0	0	0	0	0	0	86	0	0	0	0	86
θ	0	0	4	9	0	0	0	0	0	0	0	3	70	0	0	0	86
v	0	0	3	0	0	0	0	0	0	0	0	0	0	83	0	0	86
z	0	0	0	0	0	0	0	0	0	1	0	0	0	0	82	3	86
ʒ	0	0	0	0	8	0	0	0	0	0	0	0	0	1	0	77	86
total	83	85	63	91	92	86	86	86	87	84	86	92	97	91	85	82	1376

Table D.8. Confusion matrix for the consonants presented auditorily in the surgical mask condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

tape, quiet, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	85	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	86
d	2	68	6	0	8	0	0	0	0	0	0	1	1	0	0	0	86
ð	0	0	42	1	0	0	0	0	0	0	0	2	23	9	7	2	86
f	0	0	0	72	0	0	0	0	0	0	1	0	12	1	0	0	86
g	0	1	0	0	85	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	84	0	0	2	0	0	0	0	0	0	0	86
m	0	0	2	1	0	0	40	3	0	0	0	0	1	37	1	1	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	79	0	0	7	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	83	1	0	0	0	0	2	86
ʃ	0	0	0	0	1	0	0	0	0	16	64	0	1	0	0	4	86
t	0	0	0	2	0	5	0	0	8	0	0	68	2	0	1	0	86
θ	0	0	4	4	0	0	0	0	0	9	0	3	64	0	1	1	86
v	0	0	2	0	0	0	0	0	0	0	0	0	3	80	1	0	86
z	0	0	1	0	0	0	0	0	0	0	0	0	2	0	80	3	86
ʒ	0	0	0	0	5	0	0	0	0	1	0	0	0	0	30	50	86
total	87	69	57	80	99	89	40	89	89	109	66	81	109	128	121	63	1376

Table D.9. Confusion matrix for the consonants presented auditorily in the tape condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response					**balaclava (mouth hole), quiet, AV**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	83	2	0	0	0	0	0	0	0	0	0	1	0	0	0	86
ð	0	0	48	0	0	0	0	1	0	0	0	3	30	3	1	0	86
f	0	0	0	82	0	0	0	0	0	0	0	0	3	1	0	0	86
g	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	1	0	0	0	0	0	80	2	1	0	0	2	0	86
ʃ	0	0	0	0	0	0	0	0	0	1	82	0	0	0	0	3	86
t	0	0	0	0	0	0	0	0	0	0	0	82	4	0	0	0	86
θ	0	0	2	2	0	0	0	0	0	0	0	5	77	0	0	0	86
v	0	0	1	0	1	0	0	0	0	0	0	0	0	84	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	84	2	86
ʒ	0	0	0	0	6	0	0	0	0	0	0	0	0	0	1	79	86
total	86	83	53	85	93	86	86	87	86	81	84	91	115	88	88	84	1376

Table D.10. Confusion matrix for the consonants presented auditory-visually in the balaclava (mouth hole) condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response					**balaclava (no mouth hole), quiet, AV**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	80	4	0	0	0	0	0	0	0	0	0	2	0	0	0	86
ð	0	0	54	0	0	0	0	0	0	0	0	1	27	4	0	0	86
f	0	0	0	82	0	0	0	0	0	0	0	0	3	1	0	0	86
g	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	85	0	0	0	0	0	1	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	82	2	0	0	0	2	0	86
ʃ	0	0	0	0	0	0	0	0	0	2	84	0	0	0	0	0	86
t	0	0	0	0	0	0	0	0	0	0	0	86	0	0	0	0	86
θ	0	0	4	8	0	0	0	0	0	0	1	2	71	0	0	0	86
v	0	0	1	0	0	0	0	0	0	0	0	0	0	85	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	83	3	86
ʒ	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	76	86
total	86	80	64	90	95	86	85	86	86	84	87	89	104	90	85	79	1376

Table D.11. Confusion matrix for the consonants presented auditory-visually in the balaclava (no mouth hole) condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

control, quiet, AV

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	81	4	0	1	0	0	0	0	0	0	0	0	0	0	0	86
ð	0	0	52	0	0	0	0	0	0	0	0	0	33	1	0	0	86
f	0	0	0	82	0	0	0	0	0	0	0	0	2	2	0	0	86
g	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	85	0	0	0	0	1	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	85	0	0	1	0	0	0	0	86
s	0	0	1	0	0	0	0	0	0	81	2	0	0	0	2	0	86
ʃ	0	0	0	0	0	0	0	0	0	4	80	0	0	0	0	2	86
t	0	0	0	0	0	0	0	0	0	0	0	85	1	0	0	0	86
θ	0	0	2	1	0	0	0	0	0	0	0	1	82	0	0	0	86
v	0	0	0	1	0	0	0	0	0	0	0	0	0	85	0	0	86
z	0	0	0	0	0	0	0	0	0	0	1	0	0	0	82	3	86
ʒ	0	0	1	0	5	0	0	0	0	0	0	0	0	0	2	78	86
total	86	81	60	84	92	85	86	86	85	85	84	87	118	88	86	83	1376

Table D.12. Confusion matrix for the consonants presented auditory-visually in the control condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response				**hoodie/scarf combination, quiet, AV**												
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	85	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	84	2	0	0	0	0	0	0	0	0	0	0	0	0	0	86
ð	0	0	47	0	0	0	0	0	0	0	0	2	23	14	0	0	86
f	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	0	86
g	0	0	0	0	85	0	0	0	0	1	0	0	0	0	0	0	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	85	0	0	1	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	83	1	0	0	0	2	0	86
ʃ	0	0	0	0	0	0	0	0	0	1	84	0	0	0	0	1	86
t	0	0	1	0	0	0	0	0	0	0	0	83	2	0	0	0	86
θ	0	0	2	10	0	0	0	0	0	0	0	2	71	1	0	0	86
v	0	0	3	0	1	0	0	0	0	0	0	0	2	80	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	84	2	86
ʒ	0	0	3	0	7	0	0	0	0	0	0	0	0	0	1	75	86
total	85	84	58	97	93	86	86	86	85	85	85	88	98	95	87	78	1376

Table D.13. Confusion matrix for the consonants presented auditory-visually in the hoodie/scarf condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response					**motorcycle helmet, quiet, AV**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	85	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	86
d	0	81	4	0	0	0	0	0	0	0	0	0	1	0	0	0	86
ð	0	0	49	0	0	0	0	0	0	0	0	2	26	6	3	0	86
f	0	0	0	83	0	0	0	0	0	0	0	0	2	1	0	0	86
g	0	0	0	0	84	0	0	0	0	0	0	0	0	0	0	2	86
k	0	0	0	0	1	84	0	0	0	0	0	1	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	85	0	0	0	0	0	0	1	0	86
p	0	0	0	0	0	0	0	0	85	0	0	1	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	79	1	0	0	0	6	0	86
ʃ	0	0	0	0	0	0	0	0	0	1	85	0	0	0	0	0	86
t	0	0	0	0	0	0	0	0	0	0	0	82	4	0	0	0	86
θ	0	0	1	1	0	0	0	0	0	1	0	2	81	0	0	0	86
v	0	0	2	1	0	0	0	0	0	0	0	0	1	82	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	83	3	86
ʒ	0	0	1	0	7	0	0	0	0	0	1	0	0	0	1	76	86
total	85	81	57	85	92	84	86	85	85	81	87	88	115	90	94	81	1376

Table D.14. Confusion matrix for the consonants presented auditory-visually in the motorcycle helmet condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response																
	***niqāb*, quiet, AV**																
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	83	3	0	0	0	0	0	0	0	0	0	0	0	0	0	86
ð	0	0	47	0	0	0	0	0	0	0	0	1	31	5	2	0	86
f	0	0	0	74	0	0	0	0	0	1	0	0	10	1	0	0	86
g	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	2	84	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	81	1	0	0	0	4	0	86
ʃ	0	0	0	0	0	0	0	0	0	1	79	0	0	0	0	6	86
t	0	0	0	0	0	0	0	0	0	0	0	85	1	0	0	0	86
θ	0	0	1	12	0	0	0	0	0	1	0	1	71	0	0	0	86
v	0	0	1	0	0	0	0	0	0	0	0	0	0	84	1	0	86
z	0	0	0	0	0	0	0	0	0	1	0	0	0	0	83	2	86
ʒ	0	1	1	0	5	0	0	0	0	0	0	0	0	0	2	77	86
total	86	84	53	86	91	86	88	84	86	85	80	87	113	90	92	85	1376

Table D.15. Confusion matrix for the consonants presented auditory-visually in the *niqāb* condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

rubber mask, quiet, AV

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	80	1	2	0	0	0	0	0	0	0	0	0	3	0	0	0	86
d	0	77	5	0	0	0	1	0	0	0	0	0	3	0	0	0	86
ð	0	0	51	0	0	0	0	0	0	0	0	2	31	2	0	0	86
f	0	0	0	76	0	0	0	0	0	0	0	0	9	1	0	0	86
g	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	1	0	0	0	0	0	81	1	0	0	0	3	0	86
ʃ	0	0	0	0	0	1	0	0	0	4	79	0	0	0	0	2	86
t	0	0	0	0	0	0	0	0	0	0	0	84	2	0	0	0	86
θ	0	0	6	10	0	0	0	0	0	0	0	1	68	1	0	0	86
v	0	0	4	0	0	0	0	0	0	0	0	0	1	81	0	0	86
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	85	1	86
ʒ	0	0	1	0	5	0	0	0	0	0	1	0	0	0	1	78	86
total	80	78	69	87	91	87	87	86	86	85	81	87	117	85	89	81	1376

Table D.16. Confusion matrix for the consonants presented auditory-visually in the rubber mask condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response					**surgical mask, quiet, AV**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86
d	0	83	1	0	0	0	0	0	0	0	0	0	2	0	0	0	86
ð	0	1	50	0	0	0	0	0	0	0	0	1	25	8	0	1	86
f	0	0	0	80	0	0	0	0	1	0	0	0	5	0	0	0	86
g	1	0	0	0	84	0	0	1	0	0	0	0	0	0	0	0	86
k	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	86
m	0	0	0	1	0	0	85	0	0	0	0	0	0	0	0	0	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	79	1	0	0	0	5	1	86
ʃ	0	0	0	0	0	0	0	0	0	2	84	0	0	0	0	0	86
t	0	0	0	0	0	0	0	0	0	0	0	85	1	0	0	0	86
θ	0	0	2	8	0	0	0	0	0	0	0	2	74	0	0	0	86
v	0	0	1	0	0	0	0	0	0	0	0	0	0	84	1	0	86
z	0	0	0	0	0	0	0	0	0	2	0	0	1	0	77	6	86
ʒ	0	0	2	0	6	0	0	0	0	0	1	0	0	0	3	74	86
total	87	84	56	89	90	86	85	87	87	83	86	88	108	92	86	82	1376

Table D.17. Confusion matrix for the consonants presented auditory-visually in the surgical mask condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

	response					**tape, quiet, AV**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	85	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	86
d	2	69	1	0	11	0	0	0	0	0	0	0	3	0	0	0	86
ð	0	0	46	0	0	0	0	0	0	0	0	2	25	6	5	2	86
f	0	0	0	78	0	0	0	0	0	1	0	0	7	0	0	0	86
g	0	1	1	0	83	0	0	0	0	0	0	0	0	0	0	1	86
k	0	0	0	0	0	84	0	0	1	0	0	0	0	0	0	1	86
m	0	0	2	0	0	0	46	0	0	0	0	0	1	34	2	1	86
n	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	86
p	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	86
s	0	0	0	0	0	0	0	0	0	78	2	0	0	0	5	1	86
ʃ	0	0	0	0	0	0	0	0	0	19	66	0	0	0	1	0	86
t	0	0	2	2	0	1	0	0	8	0	0	70	2	0	0	1	86
θ	0	0	4	0	0	0	0	0	0	3	3	2	73	0	1	0	86
v	0	0	4	0	0	0	0	0	0	0	0	0	1	80	1	0	86
z	0	0	0	0	0	0	0	0	0	1	0	0	0	0	82	3	86
ʒ	0	0	1	0	5	0	0	0	0	0	0	0	0	0	32	48	86
total	87	70	61	80	99	85	46	86	95	102	71	74	112	121	129	58	1376

Table D.18. Confusion matrix for the consonants presented auditory-visually in the tape condition (quiet listening condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 3 (86 = 43 participants x 2 tokens per consonant).

balaclava (mouth hole), noise, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	21	1	3	18	3	0	1	0	1	0	3	0	8	12	3	4	78
d	1	27	10	0	7	1	1	7	2	4	0	5	9	2	2	0	78
ð	5	4	2	18	6	6	1	1	10	3	1	3	6	7	2	3	78
f	0	1	1	53	1	2	0	0	2	1	1	1	8	5	1	1	78
g	2	6	0	6	8	22	2	1	0	11	1	3	3	2	9	2	78
k	2	3	2	2	1	51	1	1	1	0	3	4	1	3	2	1	78
m	12	1	0	12	3	1	14	0	12	2	3	5	3	4	3	3	78
n	0	3	6	5	5	0	1	36	3	3	1	1	7	1	4	2	78
p	2	2	4	2	2	1	2	0	48	0	1	7	5	0	0	2	78
s	1	0	3	6	0	1	0	0	0	42	4	1	11	0	7	2	78
ʃ	0	0	0	0	0	0	0	0	0	1	73	1	0	0	0	3	78
t	2	4	0	4	4	13	0	1	1	1	1	36	6	3	2	0	78
θ	3	0	4	13	1	0	0	1	0	4	6	1	34	1	9	1	78
v	15	1	3	14	0	0	1	1	1	13	1	2	13	11	1	1	78
z	3	2	0	6	4	1	1	3	2	4	2	5	2	4	32	7	78
ʒ	1	6	5	7	14	0	0	4	3	1	13	1	3	3	1	16	78
total	70	61	43	166	59	99	25	56	86	90	114	76	119	58	78	48	1248

Table D.19. Confusion matrix for the consonants presented auditorily in the balaclava (mouth hole) condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

balaclava (no mouth hole), noise, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	41	3	1	3	5	11	0	1	1	2	0	0	2	5	3	0	78
d	9	7	6	1	11	8	1	10	3	1	0	0	17	1	1	2	78
ð	3	5	7	3	8	4	0	0	0	4	2	2	21	7	11	1	78
f	2	0	0	58	0	1	1	0	1	5	0	0	6	3	0	1	78
g	6	2	2	2	53	1	0	0	0	7	0	1	1	2	0	1	78
k	1	0	0	0	1	71	0	0	1	0	0	2	0	2	0	0	78
m	7	3	0	0	3	26	23	3	3	2	0	3	1	4	0	0	78
n	3	9	1	1	6	9	5	33	1	2	0	1	6	1	0	0	78
p	0	1	2	2	2	15	0	0	42	0	1	3	3	2	0	5	78
s	0	1	3	6	1	0	1	1	0	32	3	0	25	0	3	2	78
ʃ	0	0	0	0	0	0	0	0	0	3	70	0	2	0	0	3	78
t	0	0	0	1	0	13	0	0	4	0	0	56	4	0	0	0	78
θ	0	0	1	61	0	1	0	0	0	0	0	0	14	1	0	0	78
v	8	1	4	30	3	1	0	0	2	0	0	1	13	12	3	0	78
z	1	3	2	0	3	2	0	2	0	39	9	2	0	1	12	2	78
ʒ	1	2	3	1	17	5	0	5	0	1	8	0	2	1	1	31	78
total	82	37	32	169	113	168	31	55	58	98	93	71	117	42	34	48	1248

Table D.20. Confusion matrix for the consonants presented auditorily in the balaclava (no mouth hole) condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

control, noise, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	43	2	4	2	3	3	0	0	0	7	3	1	1	3	2	4	78
d	3	14	6	1	15	1	0	1	1	14	3	0	13	1	2	3	78
ð	4	6	14	4	4	2	1	1	2	1	1	1	17	14	6	0	78
f	2	0	1	62	0	0	0	2	0	1	0	0	4	6	0	0	78
g	1	3	0	1	64	1	0	0	0	5	1	0	0	0	1	1	78
k	0	0	0	2	0	74	0	0	1	1	0	0	0	0	0	0	78
m	23	2	1	5	5	6	16	1	6	2	1	2	4	3	1	0	78
n	6	9	2	3	3	7	3	20	5	2	0	4	8	5	0	1	78
p	1	1	1	2	4	23	1	0	37	1	1	4	1	1	0	0	78
s	0	0	1	5	0	0	0	2	0	57	2	0	5	0	4	2	78
ʃ	0	4	0	1	0	0	0	0	0	1	67	0	2	1	0	2	78
t	1	0	1	2	0	19	0	0	16	0	0	36	2	1	0	0	78
θ	0	1	2	26	1	0	0	0	0	3	3	1	36	4	1	0	78
v	7	2	2	34	0	1	1	0	1	0	0	0	9	21	0	0	78
z	1	5	7	4	0	2	0	1	0	1	1	1	18	3	27	7	78
ʒ	2	10	7	2	11	9	1	5	2	3	9	1	6	1	1	8	78
total	94	59	49	156	110	148	23	33	71	99	92	51	126	64	45	28	1248

Table D.21. Confusion matrix for the consonants presented auditorily in the control condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

hoodie/scarf combination, noise, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	15	7	1	3	14	9	8	1	4	7	1	1	2	0	1	4	78
d	0	23	0	4	12	9	0	8	2	2	1	4	7	1	3	2	78
ð	0	0	16	0	2	9	1	0	1	1	0	0	25	17	3	3	78
f	4	1	4	37	0	1	0	2	1	6	1	1	9	6	2	3	78
g	0	1	0	0	66	0	0	0	1	4	0	1	0	1	2	2	78
k	3	3	0	0	1	57	1	3	2	3	2	1	1	1	0	0	78
m	12	3	0	10	6	12	4	1	12	6	3	0	8	1	0	0	78
n	0	5	0	4	9	13	4	24	5	0	1	2	3	1	3	4	78
p	2	1	2	2	1	19	0	0	40	0	0	6	3	1	1	0	78
s	0	1	1	2	0	0	0	0	1	51	0	0	5	0	13	4	78
ʃ	0	0	0	0	0	0	0	0	0	1	69	0	0	0	0	8	78
t	0	0	2	0	0	12	0	0	0	0	0	59	4	1	0	0	78
θ	1	0	4	26	0	0	0	0	0	2	0	0	45	0	0	0	78
v	5	2	1	9	7	4	0	1	5	13	0	11	2	7	9	2	78
z	1	5	2	3	2	2	2	0	2	2	1	2	5	6	34	9	78
ʒ	0	3	6	2	9	0	0	1	1	5	8	1	6	1	0	35	78
total	43	55	39	102	129	147	20	41	77	103	87	89	125	44	71	76	1248

Table D.22. Confusion matrix for the consonants presented auditorily in the hoodie/scarf condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

	response																
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	6	6	5	13	8	6	3	4	2	6	4	3	6	3	1	2	78
d	3	6	0	8	7	4	6	1	2	15	5	2	9	2	4	4	78
ð	6	9	2	14	6	8	3	2	2	4	4	6	4	3	1	4	78
f	7	1	2	15	5	5	1	5	6	4	2	4	16	3	0	2	78
g	4	6	1	11	12	8	1	2	1	11	4	2	3	2	7	3	78
k	3	4	2	6	6	10	6	4	4	8	4	7	9	1	4	0	78
m	3	6	1	6	9	10	6	4	2	6	3	9	4	4	3	2	78
n	2	7	0	2	10	8	5	11	5	10	2	2	4	5	3	2	78
p	7	6	0	6	3	10	1	2	16	3	2	10	6	1	5	0	78
s	2	8	0	4	12	10	6	1	2	9	4	6	4	3	4	3	78
ʃ	1	2	0	0	0	2	0	3	2	23	28	0	8	3	1	5	78
t	8	2	2	7	5	15	1	1	15	4	1	6	9	0	0	2	78
θ	4	1	4	11	6	0	2	0	1	17	2	2	20	2	5	1	78
v	13	3	2	13	4	0	1	0	3	12	10	1	1	6	3	6	78
z	5	4	3	8	11	10	8	2	4	2	4	6	4	4	3	0	78
ʒ	4	4	2	16	10	7	2	2	1	11	3	1	4	3	4	4	78
total	78	75	26	140	114	113	52	44	68	145	82	67	111	45	48	40	1248

motorcycle helmet, noise, AO

Table D.23. Confusion matrix for the consonants presented auditorily in the motorcycle helmet condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

	response																
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	63	1	0	7	1	0	0	0	0	1	1	0	0	4	0	0	78
d	4	5	2	5	10	14	6	1	2	9	2	9	4	3	1	1	78
ð	3	3	6	5	6	3	3	7	5	9	3	1	8	6	6	4	78
f	1	3	2	47	3	3	1	1	1	0	0	0	11	5	0	0	78
g	1	0	1	0	73	1	0	0	0	0	0	0	1	0	0	1	78
k	0	1	1	2	3	61	1	0	3	1	0	0	1	1	2	1	78
m	7	4	1	4	7	21	6	5	6	6	1	3	2	3	2	0	78
n	4	8	4	4	10	10	2	6	3	6	3	3	5	4	3	3	78
p	4	1	1	1	4	5	5	2	39	2	2	1	7	3	0	1	78
s	0	0	1	0	0	1	0	1	0	54	5	0	8	0	7	1	78
ʃ	0	1	0	4	4	2	0	0	5	4	43	0	2	0	1	12	78
t	2	5	2	5	3	12	1	2	18	1	1	14	2	3	3	4	78
θ	0	2	7	30	1	0	0	0	3	0	0	1	26	5	2	1	78
v	2	3	5	16	1	10	0	2	6	5	1	4	10	8	4	1	78
z	1	0	6	4	6	9	1	2	0	3	1	2	9	9	19	6	78
ʒ	4	1	0	2	9	3	0	1	5	2	10	2	6	2	1	30	78
total	96	38	39	136	141	155	26	30	96	103	73	40	102	56	51	66	1248

***niqāb*, noise, AO**

Table D.24. Confusion matrix for the consonants presented auditorily in the *niqāb* condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

rubber mask, noise, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	26	6	3	5	4	5	2	0	3	9	1	1	5	2	3	3	78
d	6	11	4	1	1	3	0	1	1	21	11	0	8	4	3	3	78
ð	5	7	0	2	7	1	4	4	4	8	7	5	9	4	8	3	78
f	1	0	2	63	0	0	0	0	0	0	0	0	5	7	0	0	78
g	2	9	3	1	23	11	2	2	5	2	5	3	7	1	1	1	78
k	2	3	1	5	2	42	1	0	8	3	3	2	3	2	0	1	78
m	5	1	0	25	2	1	19	3	2	2	0	5	5	7	1	0	78
n	3	6	1	4	8	11	4	2	9	6	0	8	10	2	1	3	78
p	0	0	3	1	0	35	0	0	33	0	1	3	1	1	0	0	78
s	0	2	0	6	4	7	0	2	0	39	3	2	7	1	4	1	78
ʃ	0	11	3	2	2	1	0	0	0	3	40	0	3	1	1	11	78
t	2	0	3	11	1	4	0	0	14	1	0	26	11	2	0	3	78
θ	0	1	3	26	0	1	0	1	0	2	1	0	39	2	0	2	78
v	9	4	3	4	5	3	7	5	4	11	4	2	7	4	5	1	78
z	1	4	5	1	5	3	3	6	2	6	2	1	2	1	28	8	78
ʒ	1	12	4	4	17	8	0	0	2	3	2	5	8	2	6	4	78
total	63	77	38	161	81	136	42	26	87	116	80	63	130	43	61	44	1248

Table D.25. Confusion matrix for the consonants presented auditorily in the rubber mask condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

surgical mask, noise, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	21	1	1	18	9	9	2	1	2	1	0	0	7	4	1	1	78
d	1	33	4	2	12	1	0	1	0	2	0	3	13	1	2	3	78
ð	3	3	12	8	1	2	0	2	1	0	1	3	17	24	1	0	78
f	1	0	1	54	0	0	0	0	0	0	0	2	17	3	0	0	78
g	0	5	1	2	39	0	0	22	0	5	1	0	2	0	0	1	78
k	1	1	1	2	0	57	0	0	8	2	0	2	2	0	0	2	78
m	8	1	2	2	3	4	26	7	12	2	1	5	5	0	0	0	78
n	13	10	3	5	11	2	11	4	6	1	0	0	6	4	1	1	78
p	1	0	3	4	0	8	2	1	46	0	0	5	6	1	0	1	78
s	3	3	2	3	4	3	1	1	0	36	3	1	4	3	10	1	78
ʃ	0	1	0	0	1	1	0	1	0	5	67	0	1	0	0	1	78
t	0	2	4	4	6	5	0	0	2	1	0	43	8	2	0	1	78
θ	1	0	2	35	1	0	0	1	1	0	0	0	37	0	0	0	78
v	5	1	1	14	1	2	3	0	3	13	8	1	4	18	2	2	78
z	11	1	2	8	1	4	1	2	1	6	5	2	10	5	17	2	78
ʒ	3	5	2	10	9	4	0	2	4	0	3	0	4	1	1	30	78
total	72	67	41	171	98	102	46	45	86	74	89	67	143	66	35	46	1248

Table D.26. Confusion matrix for the consonants presented auditorily in the surgical mask condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

tape, noise, AO

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	5	7	2	7	20	5	4	1	7	6	4	4	2	2	2	0	78
d	1	7	2	7	13	7	3	1	1	12	8	3	3	3	4	3	78
ð	2	4	0	9	3	12	5	1	3	14	9	6	2	3	5	0	78
f	1	1	5	49	0	0	0	0	0	2	0	0	18	1	0	1	78
g	6	3	0	7	7	23	0	1	9	1	3	6	3	2	5	2	78
k	2	5	1	11	4	19	3	3	14	3	3	3	3	0	2	2	78
m	0	4	4	10	10	8	4	3	5	11	1	4	4	4	3	3	78
n	8	2	3	19	4	2	5	3	3	8	7	4	5	2	2	1	78
p	4	6	4	15	2	15	1	0	7	4	2	7	5	3	2	1	78
s	0	2	4	6	5	6	1	0	1	17	1	2	23	5	4	1	78
ʃ	3	5	4	3	5	1	0	0	1	7	18	3	12	1	7	8	78
t	2	1	1	23	4	26	3	0	6	3	1	1	2	3	2	0	78
θ	3	0	2	22	3	6	0	5	2	10	4	3	6	6	4	2	78
v	1	6	3	6	12	5	2	4	2	8	6	1	4	9	4	5	78
z	7	5	4	15	4	2	7	7	1	4	4	1	4	7	2	4	78
ʒ	5	7	3	3	8	8	1	2	5	7	4	10	6	5	3	1	78
total	50	65	42	212	104	145	39	31	67	117	75	58	102	56	51	34	1248

Table D.27. Confusion matrix for the consonants presented auditorily in the tape condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

	response					**balaclava (mouth hole), noise, AV**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	76	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	78
d	1	40	6	0	4	3	0	7	0	0	0	5	8	0	3	1	78
ð	2	3	8	2	3	2	0	1	0	0	0	9	44	2	1	1	78
f	0	0	0	68	0	0	0	0	0	0	0	0	2	8	0	0	78
g	0	4	1	1	9	26	1	2	0	13	2	8	5	1	5	0	78
k	0	0	1	1	0	67	0	0	0	0	0	6	2	0	0	1	78
m	38	0	0	0	0	0	18	0	22	0	0	0	0	0	0	0	78
n	0	13	5	1	0	2	1	39	0	3	2	5	5	0	1	1	78
p	7	0	0	1	0	0	6	0	62	0	0	0	1	0	1	0	78
s	0	0	1	4	0	0	1	0	1	53	1	1	4	0	9	3	78
ʃ	0	0	0	0	0	0	0	0	0	4	72	0	0	0	0	2	78
t	1	0	2	2	4	12	0	1	2	1	0	44	7	1	1	0	78
θ	0	0	5	1	0	0	0	0	0	8	1	1	56	1	3	2	78
v	4	0	0	35	1	0	0	0	0	0	0	0	3	34	1	0	78
z	0	3	0	3	5	3	0	2	1	4	2	2	3	5	34	11	78
ʒ	0	9	3	0	10	1	0	1	0	0	24	1	0	0	0	29	78
total	129	72	32	119	36	116	27	53	89	86	104	82	141	52	59	51	1248

Table D.28. Confusion matrix for the consonants presented auditory-visually in the balaclava (mouth hole) condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

	response					**balaclava (no mouth hole), noise, AV**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	48	3	0	3	6	4	0	0	4	1	0	0	0	6	1	2	78
d	10	19	4	1	13	4	2	6	1	0	0	0	16	1	0	1	78
ð	0	8	13	0	13	1	0	0	0	2	0	1	20	9	9	2	78
f	3	0	0	62	0	1	0	0	1	3	1	0	3	3	0	1	78
g	1	2	0	2	56	5	1	1	2	5	0	0	1	1	0	1	78
k	1	0	1	1	1	72	0	0	1	0	0	1	0	0	0	0	78
m	2	3	0	1	5	19	24	1	7	3	1	3	3	4	2	0	78
n	3	10	3	4	4	10	0	35	4	0	0	1	4	0	0	0	78
p	1	1	1	0	1	21	0	0	41	1	1	3	5	0	0	2	78
s	1	1	1	9	0	2	0	1	0	26	2	2	26	0	6	1	78
ʃ	0	0	1	1	0	0	0	0	0	1	71	0	0	0	0	4	78
t	1	0	1	1	0	12	0	0	7	0	0	53	3	0	0	0	78
θ	0	0	0	58	0	1	0	0	0	1	0	0	18	0	0	0	78
v	8	1	3	28	3	1	0	0	2	0	0	1	16	14	0	1	78
z	1	3	2	0	4	1	0	3	1	35	5	0	2	1	15	5	78
ʒ	1	4	2	0	11	5	0	5	1	1	6	1	4	1	0	36	78
total	81	55	32	171	117	159	27	52	72	79	87	66	121	40	33	56	1248

Table D.29. Confusion matrix for the consonants presented auditory-visually in the balaclava (no mouth hole) condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

control, noise, AV

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78
d	0	35	2	0	4	0	0	1	0	12	10	0	8	0	5	1	78
ð	0	0	17	0	0	0	0	0	0	0	0	1	59	0	0	1	78
f	0	0	0	75	0	0	0	0	0	0	0	0	2	1	0	0	78
g	0	2	0	0	75	0	0	0	0	0	1	0	0	0	0	0	78
k	0	0	0	0	0	77	0	0	0	0	0	0	0	0	1	0	78
m	51	1	0	0	0	0	16	0	8	1	0	0	0	1	0	0	78
n	0	16	1	0	8	6	0	28	0	0	1	15	1	0	0	2	78
p	2	0	0	0	0	0	0	0	76	0	0	0	0	0	0	0	78
s	0	0	0	0	0	0	0	0	0	70	2	0	1	0	4	1	78
ʃ	0	0	1	0	0	0	0	0	0	1	73	0	0	0	0	3	78
t	0	1	0	0	0	14	0	0	1	0	0	56	6	0	0	0	78
θ	0	0	1	1	0	0	0	0	0	0	0	1	75	0	0	0	78
v	0	0	0	49	0	0	0	0	0	1	0	0	0	27	1	0	78
z	0	0	2	0	0	0	0	1	0	1	0	1	3	0	55	15	78
ʒ	0	7	2	2	11	2	0	1	0	4	19	0	1	0	1	28	78
total	131	62	26	127	98	99	16	31	85	90	106	74	156	29	67	51	1248

Table D.30. Confusion matrix for the consonants presented auditory-visually in the control condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

	response					**hoodie/scarf combination, noise, AV**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	13	10	3	5	15	8	6	4	3	5	1	0	2	1	1	1	78
d	2	26	6	1	11	5	0	5	1	8	2	0	9	1	1	0	78
ð	0	1	15	5	2	4	0	1	0	2	0	1	28	13	3	3	78
f	5	0	2	48	0	0	0	0	2	8	0	0	11	1	1	0	78
g	0	1	0	1	68	2	0	0	0	3	1	0	0	1	1	0	78
k	1	0	1	1	2	63	0	0	1	0	0	1	1	2	2	3	78
m	15	1	0	13	3	13	6	1	9	2	1	4	2	5	3	0	78
n	1	2	4	5	4	17	1	29	1	0	1	3	5	2	2	1	78
p	1	1	1	0	1	17	1	1	43	1	0	4	2	3	1	1	78
s	0	1	0	1	0	0	0	0	0	64	2	0	2	0	5	3	78
ʃ	0	1	0	2	1	0	0	0	0	0	65	0	0	0	0	9	78
t	0	0	6	0	0	10	0	0	1	0	0	58	3	0	0	0	78
θ	0	0	4	31	0	0	0	0	0	1	0	0	41	1	0	0	78
v	1	1	6	7	1	9	1	2	10	8	0	10	2	6	9	5	78
z	2	1	3	4	3	0	2	2	1	1	0	2	2	3	44	8	78
ʒ	0	2	2	1	12	0	0	0	0	2	9	1	6	7	4	32	78
total	41	48	53	125	123	148	17	45	72	105	82	84	116	46	77	66	1248

Table D.31. Confusion matrix for the consonants presented auditory-visually in the hoodie/scarf condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

	response					**motorcycle helmet, noise, AV**											
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	11	4	4	8	4	6	2	2	13	1	3	2	6	6	4	2	78
d	4	10	3	3	2	3	3	4	4	19	5	3	9	1	2	3	78
ð	2	7	1	11	11	3	1	3	2	6	5	10	5	5	2	4	78
f	0	5	5	13	8	8	4	3	5	4	1	2	13	4	0	3	78
g	3	11	0	11	14	4	1	3	3	6	6	1	6	3	6	0	78
k	6	4	3	5	10	8	4	1	5	7	5	5	9	3	1	2	78
m	6	8	0	10	8	8	5	2	5	6	3	3	3	6	2	3	78
n	7	7	1	3	15	6	5	9	5	6	1	5	3	0	4	1	78
p	9	4	1	7	3	13	3	1	18	4	0	6	4	4	1	0	78
s	1	7	4	4	8	11	6	2	0	8	4	7	5	7	3	1	78
ʃ	0	3	0	2	1	3	0	0	1	28	27	1	3	2	4	3	78
t	4	8	1	7	2	13	0	2	12	5	1	4	12	5	0	2	78
θ	6	3	2	12	5	2	1	0	1	18	3	1	17	6	1	0	78
v	6	3	1	8	4	1	0	1	2	8	12	3	10	9	7	3	78
z	2	9	0	6	11	13	5	2	5	7	1	6	4	1	1	5	78
ʒ	5	10	2	10	4	11	3	0	0	13	4	1	11	2	1	1	78
total	72	103	28	120	110	113	43	35	81	146	81	60	120	64	39	33	1248

Table D.32. Confusion matrix for the consonants presented auditorily in the motorcycle helmet condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

	response																
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	61	2	2	5	0	0	1	0	0	0	0	2	1	4	0	0	78
d	2	6	1	9	5	20	2	3	3	8	0	5	3	5	5	1	78
ð	2	7	5	7	4	3	4	6	1	10	1	1	9	5	9	4	78
f	4	3	1	42	3	6	0	0	0	2	1	0	10	3	1	2	78
g	0	1	0	0	73	1	0	0	0	0	1	0	0	0	2	0	78
k	0	3	0	1	1	62	2	1	4	1	2	1	0	0	0	0	78
m	12	5	2	2	7	13	3	2	5	4	2	1	4	8	6	2	78
n	3	7	2	5	12	8	4	11	2	8	2	5	3	3	1	2	78
p	3	2	2	0	1	5	5	5	39	0	0	4	8	3	1	0	78
s	0	2	1	0	2	0	0	0	0	55	3	0	9	0	2	4	78
ʃ	0	1	0	4	6	0	0	0	2	4	47	1	3	0	0	10	78
t	1	4	2	2	2	12	2	1	19	2	0	19	6	2	2	2	78
θ	0	1	1	46	0	1	0	1	0	1	0	0	24	2	1	0	78
v	4	6	2	16	0	5	3	2	4	10	1	4	8	10	2	1	78
z	1	0	8	3	7	6	1	2	1	3	0	3	9	4	25	5	78
ʒ	5	3	1	7	10	2	0	0	3	3	4	0	5	3	2	30	78
total	98	53	30	149	133	144	27	34	83	111	64	46	102	52	59	63	1248

***niqāb*, noise, AV**

Table D.33. Confusion matrix for the consonants presented auditory-visually in the *niqāb* condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

	response																
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	4	6	8	0	3	2	1	0	2	9	3	3	31	1	5	0	78
d	0	14	5	0	5	1	0	1	2	20	10	1	10	1	5	3	78
ð	1	8	7	4	13	6	3	0	4	5	0	6	13	4	2	2	78
f	1	0	3	47	0	1	0	1	0	1	0	0	17	6	1	0	78
g	1	7	3	1	30	24	0	0	0	3	1	1	5	0	1	1	78
k	1	1	1	0	3	55	2	0	8	2	0	2	3	0	0	0	78
m	3	2	3	5	2	2	14	2	6	2	3	2	30	0	2	0	78
n	2	8	7	5	3	10	0	2	6	3	0	1	27	1	2	1	78
p	0	0	1	2	0	25	1	0	33	0	1	5	9	0	0	1	78
s	0	2	2	6	2	0	2	2	0	41	2	3	4	2	7	3	78
ʃ	0	9	0	0	0	0	0	0	0	7	44	0	4	0	4	10	78
t	0	0	1	2	0	12	0	0	3	4	1	39	13	0	2	1	78
θ	0	2	5	26	2	0	0	0	0	2	0	0	41	0	0	0	78
v	8	5	5	7	4	5	3	2	5	5	1	1	16	5	5	1	78
z	0	3	4	5	4	1	2	3	1	3	2	2	2	6	28	12	78
ʒ	1	14	4	0	17	6	1	2	2	3	5	4	9	0	5	5	78
total	22	81	59	110	88	150	29	15	72	110	73	70	234	26	69	40	1248

rubber mask, noise, AV

Table D.34. Confusion matrix for the consonants presented auditory-visually in the rubber mask condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

	response																
stimulus	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	30	2	1	24	3	3	0	0	3	1	1	0	4	5	0	1	78
d	1	38	6	0	12	1	0	1	0	2	0	2	11	0	1	3	78
ð	0	1	14	7	1	1	0	1	1	1	1	1	33	16	0	0	78
f	0	0	1	58	0	0	0	0	0	1	0	2	8	8	0	0	78
g	1	3	0	0	44	1	0	24	0	2	0	0	1	0	1	1	78
k	0	0	1	4	2	58	0	0	8	1	0	2	1	1	0	0	78
m	8	2	2	2	0	1	24	6	20	0	0	2	6	4	0	1	78
n	15	4	2	8	6	5	9	3	3	3	0	1	8	6	2	3	78
p	3	0	0	5	0	3	0	0	51	0	1	8	5	1	1	0	78
s	0	1	2	5	1	5	0	0	0	46	3	0	1	4	6	4	78
ʃ	0	2	0	0	3	4	0	0	0	2	64	0	2	0	0	1	78
t	3	0	3	3	1	4	0	0	4	1	0	45	10	4	0	0	78
θ	1	0	2	22	0	0	0	0	0	1	0	1	49	1	0	1	78
v	3	0	1	16	2	6	2	2	0	7	1	3	4	28	3	0	78
z	4	2	3	14	2	3	2	0	3	3	2	3	8	5	19	5	78
ʒ	4	2	4	6	18	1	1	1	2	1	2	1	6	7	3	19	78
total	73	57	42	174	95	96	38	38	95	72	75	71	157	90	36	39	1248

surgical mask, noise, AV

Table D.35. Confusion matrix for the consonants presented auditory-visually in the surgical mask condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

tape, noise, AV

stimulus \ response	b	d	ð	f	g	k	m	n	p	s	ʃ	t	θ	v	z	ʒ	total
b	44	0	0	3	3	1	4	0	19	0	1	0	1	0	0	2	78
d	1	2	2	10	9	1	2	2	1	9	14	8	4	2	6	5	78
ð	1	4	1	2	15	27	1	1	2	7	4	5	4	0	3	1	78
f	1	0	1	68	0	0	0	0	0	0	0	0	4	3	1	0	78
g	7	1	0	3	4	43	9	1	5	0	1	2	0	0	2	0	78
k	3	6	2	1	6	32	3	1	4	6	1	7	3	0	2	1	78
m	27	2	0	12	4	1	4	3	9	1	1	1	0	11	2	0	78
n	2	1	9	21	3	11	3	0	4	4	2	2	5	8	1	2	78
p	13	0	2	9	1	0	1	0	45	3	0	2	1	1	0	0	78
s	0	1	1	2	0	3	0	1	0	48	1	0	5	1	11	4	78
ʃ	0	0	2	2	8	7	1	1	1	16	19	1	7	3	6	4	78
t	1	0	0	14	1	36	0	0	7	8	1	3	3	3	1	0	78
θ	1	2	6	5	1	16	2	1	0	4	7	2	27	0	4	0	78
v	4	1	3	28	6	1	0	1	1	1	0	0	3	28	1	0	78
z	4	5	5	5	6	3	1	8	4	2	3	1	7	10	13	1	78
ʒ	2	4	1	7	14	9	2	2	2	4	3	20	3	0	3	2	78
total	111	29	35	192	81	191	33	22	104	113	58	54	77	70	56	22	1248

Table D.36. Confusion matrix for the consonants presented auditory-visually in the tape condition (speech-in-noise condition). The consonants that were presented to the participants are shown in rows, and the corresponding responses are shown in columns. Each cell contains the total count of responses for a particular stimulus. Correct responses are displayed along the diagonal, and incorrect responses above and below the diagonal. The rightmost column shows how often each consonant was presented in Experiment 4 (78 = 39 participants x 2 tokens per consonant).

D.2 *D*-prime

The following table (Table D.37) shows the results of the *d'* analysis of the perceptual consonant confusion data obtained in Experiment 3 (quiet listening condition) and Experiment 4 (speech-in-noise condition), averaged across all participants' individual *d'* results. The table contains the *d'* values and the corresponding false alarms and misses (1 – hit rates).

feature	false alarm	miss	*d'*	false alarm	miss	*d'*	false alarm	miss	*d'*	false alarm	miss	*d'*
	control, quiet, AO			**control, noise, AO**			**control, quiet, AV**			**control, noise, AV**		
plosive	0.02	0.02	4.30	0.21	0.21	1.60	0.01	0.01	4.70	0.16	0.10	2.30
fricative	0.01	0.02	4.30	0.21	0.16	1.80	0.01	0.01	4.70	0.08	0.04	3.10
nasal	0.00	0.01	5.60	0.02	0.74	1.50	0.00	0.00	6.20	0.00	0.72	2.20
bilabial	0.00	0.00	6.20	0.06	0.46	1.70	0.00	0.00	5.80	0.00	0.01	5.30
labiodental	0.01	0.03	4.10	0.09	0.21	2.10	0.00	0.01	5.20	0.00	0.03	4.60
dental	0.01	0.13	3.40	0.10	0.56	1.20	0.01	0.02	4.60	0.03	0.03	3.90
alveolar	0.00	0.04	4.30	0.10	0.49	1.30	0.01	0.04	4.20	0.03	0.23	2.70
postalveolar	0.01	0.05	4.20	0.03	0.45	2.00	0.01	0.07	4.00	0.03	0.21	2.70
velar	0.01	0.01	4.60	0.11	0.11	2.50	0.01	0.01	4.90	0.04	0.03	3.70
voiced	0.00	0.04	4.40	0.09	0.35	1.70	0.01	0.05	3.90	0.03	0.29	2.50
voiceless	0.04	0.00	4.40	0.35	0.09	1.70	0.05	0.01	3.90	0.29	0.03	2.50
	balaclava (mouth hole), quiet, AO			**balaclava (mouth hole), noise, AO**			**balaclava (mouth hole), quiet, AV**			**balaclava (mouth hole), noise, AV**		
plosive	0.01	0.01	4.40	0.20	0.37	1.16	0.02	0.01	4.30	0.18	0.18	1.80
fricative	0.01	0.02	4.50	0.35	0.20	1.24	0.01	0.02	4.30	0.14	0.11	2.30
nasal	0.00	0.00	6.00	0.03	0.67	1.47	0.00	0.00	6.20	0.02	0.63	1.70
bilabial	0.00	0.00	5.60	0.07	0.52	1.46	0.00	0.00	6.20	0.02	0.02	4.30
labiodental	0.01	0.03	4.20	0.13	0.47	1.21	0.00	0.03	4.50	0.02	0.07	3.50
dental	0.01	0.10	3.70	0.11	0.71	0.71	0.01	0.09	3.70	0.06	0.28	2.20
alveolar	0.01	0.03	4.40	0.15	0.41	1.28	0.01	0.03	4.20	0.09	0.31	1.80
postalveolar	0.00	0.08	4.40	0.05	0.33	2.07	0.00	0.05	4.40	0.03	0.19	2.80
velar	0.01	0.00	5.40	0.07	0.47	1.54	0.01	0.00	5.60	0.05	0.35	2.10
voiced	0.01	0.05	4.20	0.17	0.42	1.16	0.01	0.04	4.00	0.11	0.36	1.60
voiceless	0.05	0.01	4.20	0.42	0.17	1.16	0.04	0.01	4.00	0.36	0.11	1.60
	tape, quiet, AO			**tape, noise, AO**			**tape, quiet, AV**			**tape, noise, AV**		
plosive	0.01	0.03	4.20	0.29	0.44	0.70	0.01	0.03	4.20	0.32	0.32	0.93
fricative	0.08	0.01	3.60	0.44	0.33	0.59	0.08	0.01	3.70	0.33	0.33	0.87
nasal	0.00	0.26	3.70	0.05	0.90	0.34	0.00	0.24	3.80	0.04	0.94	0.22
bilabial	0.01	0.21	3.10	0.12	0.84	0.19	0.01	0.16	3.30	0.08	0.29	1.95
labiodental	0.04	0.11	2.90	0.19	0.58	0.68	0.03	0.08	3.20	0.12	0.19	2.05
dental	0.03	0.23	2.70	0.12	0.94	-0.36	0.02	0.14	3.10	0.07	0.76	0.80
alveolar	0.08	0.10	2.70	0.26	0.75	-0.04	0.07	0.09	2.80	0.16	0.65	0.61
postalveolar	0.01	0.31	2.90	0.07	0.80	0.62	0.01	0.34	2.70	0.05	0.82	0.75
velar	0.01	0.02	4.30	0.18	0.66	0.50	0.01	0.03	4.10	0.17	0.46	1.06
voiced	0.02	0.05	3.70	0.28	0.54	0.50	0.02	0.04	3.70	0.21	0.51	0.78
voiceless	0.05	0.02	3.70	0.54	0.27	0.50	0.04	0.02	3.70	0.51	0.21	0.78

(table continues on next page)

feature	false alarm	miss	*d'*	false alarm	miss	*d'*	false alarm	miss	*d'*	false alarm	miss	*d'*
	surgical mask, quiet, AO			**surgical mask, noise, AO**			**surgical mask, quiet, AV**			**surgical mask, noise, AV**		
plosive	0.02	0.01	4.40	0.22	0.31	1.30	0.01	0.01	4.60	0.19	0.28	1.40
fricative	0.01	0.02	4.40	0.24	0.17	1.70	0.01	0.01	4.60	0.25	0.15	1.70
nasal	0.00	0.01	5.70	0.04	0.69	1.30	0.00	0.01	5.70	0.03	0.73	1.20
bilabial	0.00	0.01	5.40	0.08	0.49	1.40	0.00	0.00	5.60	0.07	0.41	1.70
labiodental	0.01	0.04	4.00	0.14	0.43	1.30	0.01	0.05	3.90	0.14	0.29	1.60
dental	0.01	0.12	3.60	0.11	0.56	1.10	0.01	0.12	3.50	0.09	0.37	1.70
alveolar	0.01	0.02	4.40	0.12	0.53	1.10	0.01	0.03	4.10	0.11	0.53	1.20
postalveolar	0.00	0.06	4.20	0.03	0.35	2.20	0.01	0.08	3.90	0.03	0.45	2.10
velar	0.01	0.01	4.70	0.10	0.38	1.60	0.01	0.01	4.80	0.08	0.33	1.90
voiced	0.01	0.04	4.00	0.13	0.37	1.50	0.01	0.04	4.00	0.11	0.36	1.60
voiceless	0.04	0.01	4.00	0.37	0.13	1.50	0.04	0.01	4.00	0.36	0.11	1.60
	balaclava (no mouth hole), quiet, AO			**balaclava (no mouth hole), noise, AO**			**balaclava (no mouth hole), quiet, AV**			**balaclava (no mouth hole), noise, AV**		
plosive	0.01	0.03	4.00	0.20	0.20	1.70	0.01	0.01	4.40	0.20	0.16	1.86
fricative	0.03	0.02	4.10	0.16	0.15	2.00	0.01	0.02	4.50	0.14	0.15	2.12
nasal	0.00	0.00	6.20	0.02	0.59	1.80	0.00	0.01	5.70	0.02	0.62	1.82
bilabial	0.00	0.01	5.40	0.05	0.50	1.60	0.00	0.01	5.50	0.05	0.46	1.73
labiodental	0.01	0.03	4.10	0.10	0.34	1.70	0.01	0.02	4.30	0.10	0.31	1.79
dental	0.02	0.11	3.30	0.10	0.72	0.70	0.01	0.09	3.60	0.09	0.67	0.87
alveolar	0.01	0.06	3.90	0.09	0.45	1.40	0.00	0.03	4.50	0.08	0.45	1.54
postalveolar	0.01	0.06	4.10	0.03	0.28	2.50	0.00	0.07	4.10	0.02	0.25	2.66
velar	0.01	0.01	4.70	0.14	0.19	1.90	0.01	0.00	5.50	0.13	0.14	2.20
voiced	0.01	0.05	4.20	0.07	0.38	1.80	0.01	0.04	4.00	0.06	0.35	1.93
voiceless	0.05	0.01	4.20	0.38	0.07	1.80	0.04	0.01	4.00	0.35	0.06	1.93
	hoodie/scarf combination, quiet, AO			**hoodie/scarf combination, noise, AO**			**hoodie/scarf combination, quiet, AV**			**hoodie/scarf combination, noise, AV**		
plosive	0.01	0.01	4.70	0.21	0.20	1.70	0.01	0.01	4.40	0.19	0.21	1.70
fricative	0.01	0.01	4.70	0.18	0.15	1.90	0.01	0.02	4.40	0.21	0.13	1.90
nasal	0.00	0.00	6.20	0.03	0.79	1.10	0.00	0.00	6.20	0.02	0.76	1.30
bilabial	0.00	0.00	6.20	0.04	0.59	1.50	0.00	0.01	5.60	0.03	0.59	1.60
labiodental	0.02	0.05	3.70	0.08	0.62	1.10	0.02	0.03	3.90	0.10	0.60	1.00
dental	0.01	0.17	3.30	0.07	0.42	1.70	0.01	0.17	3.30	0.07	0.44	1.60
alveolar	0.01	0.03	4.40	0.14	0.38	1.40	0.01	0.02	4.40	0.12	0.35	1.60
postalveolar	0.00	0.06	4.20	0.04	0.23	2.50	0.00	0.07	4.30	0.03	0.26	2.50
velar	0.01	0.01	5.00	0.14	0.21	1.90	0.01	0.01	4.90	0.13	0.13	2.30
voiced	0.02	0.03	4.00	0.14	0.37	1.40	0.01	0.04	4.10	0.11	0.35	1.60
voiceless	0.03	0.02	4.00	0.37	0.14	1.40	0.04	0.01	4.10	0.35	0.11	1.60

(table continues on next page)

feature	false alarm	miss	*d'*	false alarm	miss	*d'*	false alarm	miss	*d'*	false alarm	miss	*d'*
	***niqāb*, quiet, AO**			***niqāb*, noise, AO**			***niqāb*, quiet, AV**			***niqāb*, noise, AV**		
plosive	0.01	0.01	4.60	0.26	0.23	1.37	0.01	0.01	4.70	0.25	0.22	1.44
fricative	0.01	0.01	4.70	0.23	0.22	1.52	0.01	0.01	4.80	0.22	0.21	1.56
nasal	0.00	0.00	6.20	0.03	0.88	0.66	0.00	0.00	6.00	0.04	0.87	0.64
bilabial	0.00	0.00	6.00	0.09	0.44	1.50	0.00	0.00	6.00	0.08	0.45	1.55
labiodental	0.01	0.08	3.60	0.11	0.51	1.21	0.01	0.08	3.70	0.12	0.54	1.07
dental	0.01	0.14	3.30	0.09	0.70	0.84	0.01	0.13	3.40	0.09	0.75	0.70
alveolar	0.01	0.03	4.10	0.11	0.58	1.00	0.01	0.03	4.30	0.15	0.54	0.96
postalveolar	0.01	0.05	4.20	0.04	0.39	2.02	0.00	0.06	4.40	0.03	0.42	2.05
velar	0.00	0.00	5.80	0.15	0.12	2.26	0.01	0.00	5.00	0.13	0.12	2.30
voiced	0.02	0.05	3.80	0.22	0.39	1.05	0.02	0.04	3.80	0.18	0.36	1.28
voiceless	0.05	0.02	3.80	0.40	0.22	1.05	0.04	0.02	3.80	0.36	0.18	1.28
	rubber mask, quiet, AO			**rubber mask, noise, AO**			**rubber mask, quiet, AV**			**rubber mask, noise, AV**		
plosive	0.01	0.02	4.30	0.27	0.37	0.96	0.01	0.03	4.20	0.24	0.37	1.02
fricative	0.01	0.02	4.40	0.37	0.29	0.89	0.02	0.01	4.30	0.42	0.26	0.84
nasal	0.00	0.00	6.20	0.04	0.82	0.87	0.00	0.00	6.20	0.02	0.88	0.78
bilabial	0.00	0.00	6.20	0.10	0.62	0.98	0.00	0.02	5.10	0.06	0.73	0.97
labiodental	0.01	0.06	3.80	0.12	0.50	1.20	0.01	0.08	3.70	0.07	0.58	1.30
dental	0.01	0.12	3.40	0.11	0.67	0.79	0.02	0.09	3.30	0.21	0.58	0.62
alveolar	0.01	0.03	4.30	0.19	0.54	0.78	0.01	0.03	4.20	0.17	0.50	0.95
postalveolar	0.01	0.07	4.00	0.06	0.63	1.20	0.00	0.07	4.40	0.05	0.59	1.47
velar	0.01	0.01	5.00	0.13	0.50	1.14	0.00	0.00	5.70	0.12	0.28	1.78
voiced	0.02	0.05	3.70	0.16	0.45	1.11	0.02	0.05	3.70	0.15	0.50	1.03
voiceless	0.05	0.02	3.70	0.45	0.16	1.11	0.05	0.02	3.70	0.50	0.15	1.03
	motorcycle helmet, quiet, AO			**motorcycle helmet, noise, AO**			**motorcycle helmet, quiet, AV**			**motorcycle helmet, noise, AV**		
plosive	0.01	0.01	4.70	0.37	0.52	0.28	0.01	0.02	4.30	0.39	0.50	0.29
fricative	0.00	0.02	4.70	0.43	0.41	0.41	0.02	0.02	4.20	0.41	0.40	0.47
nasal	0.00	0.00	6.20	0.06	0.83	0.55	0.00	0.01	5.70	0.05	0.87	0.52
bilabial	0.00	0.00	6.20	0.15	0.80	0.18	0.00	0.01	5.60	0.12	0.69	0.66
labiodental	0.01	0.04	4.00	0.14	0.76	0.39	0.01	0.03	4.40	0.14	0.78	0.31
dental	0.01	0.13	3.50	0.10	0.81	0.42	0.01	0.09	3.60	0.11	0.84	0.22
alveolar	0.01	0.02	4.40	0.30	0.69	0.01	0.01	0.03	4.10	0.28	0.64	0.21
postalveolar	0.00	0.06	4.30	0.08	0.74	0.78	0.00	0.06	4.20	0.07	0.78	0.70
velar	0.01	0.00	5.50	0.18	0.77	0.20	0.01	0.02	4.60	0.17	0.77	0.21
voiced	0.01	0.03	4.20	0.33	0.51	0.40	0.01	0.04	3.90	0.35	0.52	0.32
voiceless	0.03	0.01	4.20	0.51	0.33	0.40	0.04	0.01	3.90	0.52	0.35	0.32

D.3 Facewear effects within modalities

In §5.5.2, participants' consonant identification performance in the AO condition was compared to their performance in the AV condition. The *d'* values obtained in the AO condition for each feature were therefore compared with the corresponding *d'* values elicited in the AV condition *within* each facewear condition (a 'vertical' comparison regarding Figures 5.12 to 5.21 presented in §5.5.2). We can also evaluate the extent to which facewear changed the perceivers' sensitivity to phonetic features within a given listening condition and modality. To do so, *d'* values *between* facewear conditions were statistically compared (a 'horizontal' comparison regarding Figures 5.12 to 5.21). It was most feasible to contrast each facewear condition with the control condition only. The results of *post-hoc* Bonferroni-adjusted pairwise comparisons are shown in Table D.38. To emphasise once more, the table does *not* show how well a phonetic feature was detected in a particular facewear condition, but whether *d'* obtained in each of the facewear conditions significantly differed from *d'* obtained in the control condition.

In the 'quiet' condition (upper half of Table D.38), most levels of comparison were non-significant ('ns'). This means that sensitivity to most features differed only marginally, or not at all, when the consonants (bearing their respective features) were spoken through facewear, or when they were produced without the talker wearing facewear. The only exception here is the tape condition, where many levels of comparison were highly significant. Sensitivity to most features was reduced in the tape condition, i.e., participants detected most features less reliably.

In the noisy condition (lower half of Table D.38), most comparisons between control and facewear achieved significance. By and large, sensitivity to phonetic features improved more markedly in the AV than in the AO modality across facewear conditions, especially in the surgical mask, balaclava (no mouth hole), and hoodie/scarf conditions.

Based on these findings it can be argued that the perceptual (auditory) effect of the facewear-induced acoustic changes to the speech signal *alone* (AO condition) appears to be less prominent than the perceptual (auditory-visual) effect caused by

the acoustic changes and the deficit of visual speech cues brought about by facial occlusion (AV condition). Put another way, losing visual information caused more of a problem perceptually than did the loss of auditory information.

difference from control		balaclava 2	tape	surgical mask	balaclava 1	hoodie/ scarf	*niqāb*	rubber mask	helmet
quiet listening condition (Experiment 3)									
plosive	**AO**	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
	AV	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
fricative	**AO**	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
	AV	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
nasal	**AO**	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
	AV	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
bilabial	**AO**	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
	AV	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
labiodental	**AO**	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
	AV	*ns*	***	*	*ns*	*ns*	**	**	*ns*
dental	**AO**	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
	AV	*ns*	**	*	*ns*	*	**	**	*ns*
alveolar	**AO**	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
	AV	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
postalveolar	**AO**	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
	AV	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
velar	**AO**	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
	AV	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
voicing	**AO**	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*	*ns*
	AV	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
speech-in-noise (Experiment 4)									
plosive	**AO**	**	***	*ns*	*ns*	*ns*	*ns*	**	***
	AV	**	***	***	*ns*	*	***	***	***
fricative	**AO**	***	***	*ns*	*ns*	*ns*	*ns*	***	***
	AV	***	***	***	***	***	***	***	***
nasal	**AO**	*ns*	***	*ns*	*ns*	*ns*	**	*ns*	**
	AV	*ns*	***	*	*ns*	*ns*	***	***	***
bilabial	**AO**	*ns*	***	*ns*	*ns*	*ns*	*ns*	***	***
	AV	*ns*	***	***	***	***	***	***	***
labiodental	**AO**	**	***	***	***	***	***	***	***
	AV	*ns*	***	***	***	***	***	***	***
dental	**AO**	*ns*	***	*ns*	**	**	*ns*	*ns*	***
	AV	***	***	***	***	***	***	***	***
alveolar	**AO**	*ns*	***	*ns*	*ns*	*ns*	*	***	***
	AV	***	***	***	***	***	***	***	***
postalveolar	**AO**	*ns*	***	*ns*	***	***	*ns*	**	***
	AV	*ns*	***	**	*ns*	*ns*	**	***	***
velar	**AO**	***	***	*ns*	*ns*	*ns*	*ns*	***	***
	AV	***	***	*ns*	***	***	***	***	***
voicing	**AO**	**	***	*ns*	*ns*	*ns*	***	***	***
	AV	*	***	***	***	***	***	***	***

Table D.38. Differences between the *d'* values obtained in each facewear condition and the corresponding *d'* values in the control condition (Experiments 3 and 4). '***' denotes a significant difference at $p < .001$, '**' at $p < .01$, and '*' at $p < .05$, and '*ns*' indicates non-significance.

D.4 Effect of order

The possible occurrence of a response bias on the part of the listeners, i.e., the listeners' tendency to favour the first or second pair response across experimental trials, was controlled for in Experiment 5 (see Chapter 6). This was done by counterbalancing the presentation order of pairs. In 'order 1', the same-talker pair was presented first, and the different-talker pair second. In 'order 2', this sequence of pairs was swapped. Pair 1 now consisted of the speech tokens of the different talkers, and pair 2 contained the speech tokens of the same talker.

ANOVA revealed a significant main effect of order on response accuracy [$F(1,23) = 9.55$, $p < .01$, $\eta_p^2 = .29$], which is indicative of an overall weak response bias. The interaction between order and facewear was also found to be significant [$F(2,46) = 7.40$, $p < .01$, $\eta_p^2 = .24$]. When rerunning ANOVAs for each facewear condition separately, no evidence of a response bias was found for the control trials. This is illustrated by the near-horizontal line in the left graph in Figure D.1.

The helmet [$F(1,23) = 9.78$, $p < .01$, $\eta_p^2 = .30$] and tape [$F(1,23) = 9.96$, $p < .01$, $\eta_p^2 = .30$], on the other hand, each produced a significant bias. This means that the listeners performed significantly better at discriminating between the talkers when the same-talker pairs were presented prior to the different-talker pairs, than vice versa (see the drop in the percentage correct scores for order 2 in the left graph in Figure D.1). *Post-hoc* Bonferroni-adjusted pairwise comparisons revealed that the number of correct talker discriminations in the helmet condition (order 1 = 78.9%, order 2 = 69.6%) and in the tape condition (order 1 = 71.8%, order 2 = 63.4%) was significantly lower in order 2 than in order 1 (ps < .01), respectively.

Lastly, as indicated in the right graph in Figure D.1, the mean response times significantly increased in order 2 [$F(1,23) = 7.65$, $p < .05$, $\eta_p^2 = .25$] for all facewear conditions, namely by around 10% in the control, 7% in the helmet, and 6% in the tape condition.

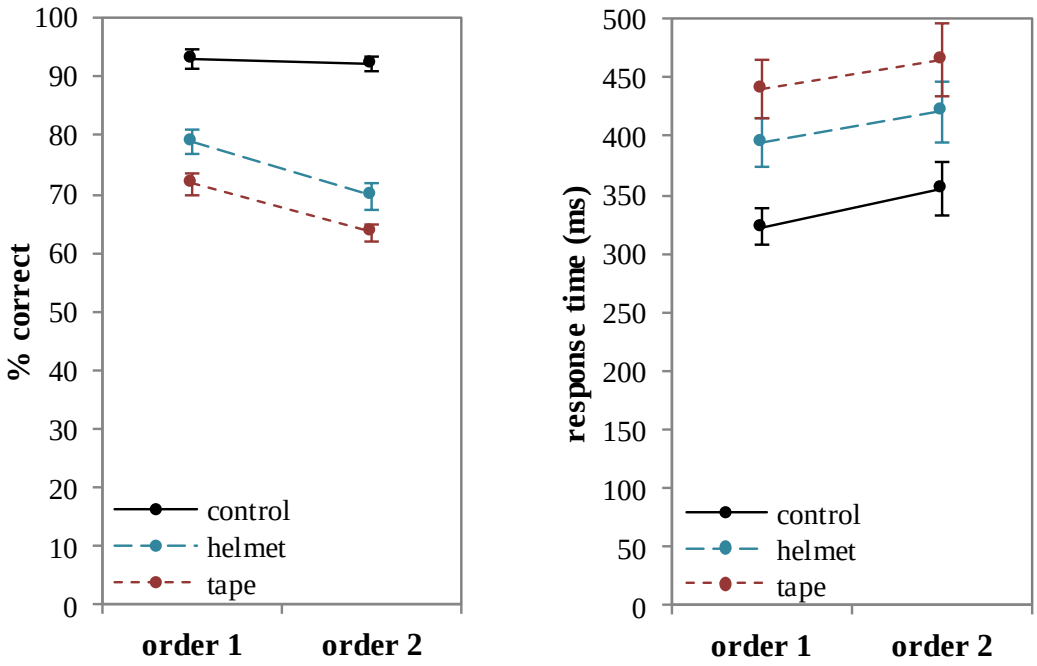

Figure D.1. *Left*: Response accuracy as a function of facewear for both presentation orders (order 1 = same-talker pair + different-talker pair; order 2 = different-talker pair + same-talker pair). Talker discrimination accuracy significantly dropped in the helmet and tape conditions (*p*s < .01), which is evidence for a response bias. No response bias was found in the control condition. *Right*: Reaction time (in ms) as a function of facewear for both orders. In the control and helmet conditions, the listeners were significantly slower in responding to the trials in which the different-talker pair was played first (*p*s < .05). The error bars show the standard error of the mean.

D.5 ANOVAs

fricatives (Experiment 1)						
within-subject factor	***F*- value / *df*s**	***p*- value**	η_p^2	***F*- value / *df*s**	***p*- value**	η_p^2
	intensity			**spectral peak**		
fricative	$F(3,15) = 78.10$	**.000 *****	.94	$F(3,15) = 10.99$	**.000 *****	.69
facewear	$F(7,35) = 10.88$	**.000 *****	.69	$F(7,35) = .93$	**.500 *ns***	.16
syllable	$F(1,5) = 47.96$	**.001 ****	.91	$F(1,5) = 38.45$	**.002 ****	.89
fricative * facewear	$F(21,105) = 2.70$	**.000 *****	.35	$F(21,105) = 1.66$	**.050 ***	.25
fricative * syllable	$F(3,15) = 27.23$	**.000 *****	.85	$F(3,15) = 5.19$	**.012 ***	.51
facewear * syllable	$F(7,35) = 2.65$	**.026 ***	.35	$F(7,35) = 1.26$	**.297 *ns***	.20
fricative * facewear * syllable	$F(21,105) = 1.61$	**.060 *ns***	.24	$F(21,105) = 1.25$	**.227 *ns***	.20
	centre of gravity			**standard deviation**		
fricative	$F(3,15) = 49.60$	**.000 *****	.91	$F(3,15) = 54.33$	**.000 *****	.92
facewear	$F(7,35) = 8.44$	**.000 *****	.63	$F(7,35) = .81$	**.583 *ns***	.14
syllable	$F(1,5) = 122.65$	**.000 *****	.96	$F(1,5) = 2.19$	**.199 *ns***	.30
fricative * facewear	$F(21,105) = 3.94$	**.000 *****	.44	$F(21,105) = 2.35$	**.002 ****	.32
fricative * syllable	$F(3,15) = 27.84$	**.000 *****	.85	$F(3,15) = 2.90$	**.070 *ns***	.37
facewear * syllable	$F(7,35) = .57$	**.775 *ns***	.10	$F(7,35) = 2.03$	**.079 *ns***	.29
fricative * facewear * syllable	$F(21,105) = 1.10$	**.406 *ns***	.17	$F(21,105) = 2.28$	**.003 ****	.30
	skewness			**kurtosis**		
fricative	$F(3,15) = 62.36$	**.000 *****	.93	$F(3,15) = 38.30$	**.000 *****	.89
facewear	$F(7,35) = 2.61$	**.028 ***	.34	$F(7,35) = 3.00$	**.014 ***	.38
syllable	$F(1,5) = 15.25$	**.011 ***	.75	$F(1,5) = 2.21$	**.197 *ns***	.31
fricative * facewear	$F(21,105) = 2.67$	**.001 ****	.35	$F(21,105) = 2.20$	**.005 ****	.31
fricative * syllable	$F(1,3) = 6.6$[a]	**.027 ***	.57	$F(3,15) = 3.50$	**.042 ***	.41
facewear * syllable	$F(7,35) = .32$	**.940 *ns***	.06	$F(7,35) = .22$	**.987 *ns***	.04
fricative * facewear * syllable	$F(21,105) = 1.08$	**.380 *ns***	.18	$F(21,105) = .53$	**.950 *ns***	.10

[a] $X^2(5) = 13.63$, $p < .05$, $\varepsilon = .50$; ε = Greenhouse-Geisser correction; *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.39. Summary of results of three-way repeated-measures ANOVAs for each dependent variable separately (intensity, spectral peak, centre of gravity, standard deviation, skewness, kurtosis), as a function of fricative, facewear, and syllable position.

fricatives (Experiment 1)						
within-subject factor	***F*- value / *df*s**	***p*- value**		η_p^2	***F*- value / *df*s**	***p*- value**		η_p^2
	intensity				**spectral peak**			
fricative	$F(3,18) = 1824.60$	**.000**	***	.99	$F(3,18) = 37.07$	**.000**	***	.86
facewear	$F(7,42) = 346.62$	**.000**	***	.98	$F(7,42) = 7.10$	**.000**	***	.54
fricative * facewear	$F(21,126) = 6.73$	**.000**	***	.53	$F(21,126) = 7.95$	**.000**	***	.57
	centre of gravity				**standard deviation**			
fricative	$F(3,18) = 472.48$	**.000**	***	.99	$F(3,18) = 626.51$	**.000**	***	.99
facewear	$F(7,42) = 25.94$	**.000**	***	.81	$F(7,42) = 1.05$	**.412**	***ns***	.15
fricative * facewear	$F(21,126) = 8.63$	**.000**	***	.59	$F(21,126) = 3.53$	**.000**	***	.37
	skewness				**kurtosis**			
fricative	$F(3,18) = 242.87$	**.000**	***	.98	$F(2,11) = 82.32$[a]	**.000**	***	.93
facewear	$F(7,42) = 4.01$	**.002**	**	.40	$F(7,42) = 1.63$	**.153**	***ns***	.21
fricative * facewear	$F(21,126) = 6.62$	**.000**	***	.53	$F(21,126) = 1.68$	**.042**	*	.22

[a] $X^2(5) = 12.15$, $p < .05$, $\varepsilon = .62$; ε = Greenhouse-Geisser; *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.40. Summary of results of two-way repeated-measures ANOVAs for each dependent variable separately (intensity, spectral peak, centre of gravity, standard deviation, skewness, kurtosis), as a function of fricative and facewear (syllable onset data only).

quiet listening condition (Experiment 3)			
within-subject factor	***F*- value / *df*s**	***p*- value**		η_p^2
modality	$F(1,42) = 5.11$	**.029**	*	.11
facewear	$F(6,239) = 87.43$[a]	**.000**	***	.68
consonant	$F(3,120) = 26.90$[b]	**.000**	***	.39
modality * facewear	$F(6,248) = 1.48$[c]	**.187**	***ns***	.03
modality * consonant	$F(7,304) = 2.01$[d]	**.052**	***ns***	.05
facewear * consonant	$F(120,5040) = 11.37$	**.000**	***	.21
modality * facewear * consonant	$F(120,5040) = 1.23$	**.048**	*	.03

[a] $X^2(35) = 61.54$, $p < .01$, $\varepsilon = .71$; [b] $X^2(119) = 1005.96$, $p < .001$, $\varepsilon = .19$

[c] $X^2(35) = 65.49$, $p < .001$, $\varepsilon = .74$; [d] $X^2(119) = 324.61$, $p < .001$, $\varepsilon = .48$

ε = Greenhouse-Geisser correction, *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.41. Summary of results of a three-way repeated-measures ANOVA with percentage correct consonant identification as the dependent variable, as a function of modality, facewear and consonant (Experiment 3).

quiet listening condition (Experiment 3)						
within-subject factor	***F*- value / *df*s**	***p*- value**	η_p^2	***F*- value / *df*s**	***p*- value**	η_p^2
	auditory-only (AO)			**auditory-visual (AV)**		
facewear	$F(6,245) = 56.20$[a]	**.000 *****	.57	$F(6,246) = 38.28$[c]	**.000 *****	.48
consonant	$F(3,134) = 27.65$[b]	**.000 *****	.40	$F(3,119) = 23.78$[d]	**.000 *****	.36
facewear * consonant	$F(120,5040) = 7.67$	**.000 *****	.15	$F(120,5040) = 6.27$	**.000 *****	.13

[a] $X^2(35) = 63.89$, $p < .01$, $\varepsilon = .73$; [b] $X^2(119) = 852.65$, $p < .001$, $\varepsilon = .21$; [c] $X^2(35) = 71.83$, $p < .001$, $\varepsilon = .73$
[d] $X^2(119) = 963.10$, $p < .001$, $\varepsilon = .19$; ε = Greenhouse-Geisser, *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.42. Summary of results of two-way repeated-measures ANOVAs with percentage correct consonant identification as the dependent variable, for each modality (AO/AV) separately, as a function of facewear and consonant (Experiment 3).

quiet listening condition (Experiment 3)						
within-subject factor	***F*- value / *df*s**	***p*- value**	η_p^2	***F*- value / *df*s**	***p*- value**	η_p^2
	control					
modality	$F(1,42) = 1.55$	**.220 *ns***	.04			
consonant	$F(3,132) = 19.50$[a]	**.000 *****	.32			
modality * consonant	$F(8,328) = 1.94$[b]	**.055 *ns***	.04			
	balaclava (mouth hole)			**balaclava (no mouth hole)**		
modality	$F(1,42) = .83$	**.368 *ns***	.02	$F(1,42) = 3.12$	**.084 *ns***	.07
consonant	$F(15,630) = 24.46$	**.000 *****	.37	$F(15,630) = 14.12$	**.000 *****	.25
modality * consonant	$F(15,630) = .69$	**.790 *ns***	.02	$F(15,630) = 1.43$	**.130 *ns***	.03
	motorcycle helmet			**hoodie/scarf combination**		
modality	$F(1,42) = .31$	**.583 *ns***	.01	$F(1,42) = .59$	**.449 *ns***	.01
consonant	$F(4,165) = 22.32$[c]	**.000 *****	.35	$F(15,630) = 20.81$	**.000 *****	.33
modality * consonant	$F(7,289) = .84$[d]	**.552 *ns***	.02	$F(15,630) = 1.15$	**.338 *ns***	.03
	niqāb			**rubber mask**		
modality	$F(1,42) = .11$	**.737 *ns***	.00	$F(1,42) = .10$	**.757 *ns***	.00
consonant	$F(15,630) = 21.20$	**.000 *****	.34	$F(15,630) = 20.38$	**.000 *****	.33
modality * consonant	$F(15,630) = 1.27$	**.214 *ns***	.03	$F(15,630) = 1.55$	**.084 *ns***	.04
	surgical mask			**tape**		
modality	$F(1,42) = .86$	**.358 *ns***	.02	$F(1,42) = 6.45$	**.015 ***	.13
consonant	$F(15,630) = 14.84$	**.000 *****	.27	$F(6,262) = 32.38$[e]	**.000 *****	.44
modality * consonant	$F(15,630) = 1.06$	**.391 *ns***	.03	$F(8,341) = 1.69$[f]	**.049 ***	.04

[a] $X^2(119) = 734.55$, $p < .001$, $\varepsilon = .21$; [b] $X^2(119) = 392.16$, $p < .001$, $\varepsilon = .52$; [c] $X^2(119) = 718.96$, $p < .001$, $\varepsilon = .26$
[d] $X^2(119) = 468.97$, $p < .001$, $\varepsilon = .46$; [e] $X^2(119) = 389.32$, $p < .001$, $\varepsilon = .42$; [f] $X^2(119) = 317.92$, $p < .001$, $\varepsilon = .54$
ε = Greenhouse-Geisser correction, *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.43. Summary of results of two-way repeated-measures ANOVAs with percentage correct consonant identification as the dependent variable, for control and each facewear condition separately, as a function of modality and consonant (Experiment 3).

speech-in-noise (Experiment 4)			
within-subject factor	***F*- value / *df*s**	***p*- value**	η_p^2
modality	F (1,38) = 196.12	**.000 *****	.84
facewear	F (5,207) = 291.93[a]	**.000 *****	.89
consonant	F (10,378) = 105.96[b]	**.000 *****	.74
modality * facewear	F (8,304) = 37.13	**.000 *****	.49
modality * consonant	F (10,368) = 7.70[c]	**.000 *****	.17
facewear * consonant	F (120,4560) = 24.01	**.000 *****	.39
modality * facewear * consonant	F (120,4560) = 4.81	**.000 *****	.11

[a] X^2 (35) = 62.09, $p < .01$, $\varepsilon = .68$; [b] X^2 (119) = 158.54, $p < .05$, $\varepsilon = .66$

[c] X^2 (119) = 173.13, $p < .01$, ε = .65; ε = Greenhouse-Geisser correction

*** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.44. Summary of results of a three-way repeated-measures ANOVA with percentage correct consonant identification as the dependent variable, as a function of modality, facewear and consonant (Experiment 4).

speech-in-noise (Experiment 4)						
within-subject factor	***F*- value / *df*s**	***p*- value**	η_p^2	***F*- value / *df*s**	***p*- value**	η_p^2
	auditory-only (AO)			**auditory-visual (AV)**		
facewear	F (7,213) = 145.85[a]	**.000 *****	.79	F (8,304) = 262.86	**.000 *****	.87
consonant	F (15,570) = 80.30	**.000 *****	.68	F (10,370) = 94.68[b]	**.000 *****	.71
facewear * consonant	F (120,4560) = 14.65	**.000 *****	.28	F (120,4560) = 18.06	**.000 *****	.32

[a] X^2 (35) = 60.41, $p < .01$, $\varepsilon = .70$; [b] X^2 (119) = 168.32, $p < .01$, $\varepsilon = .65$; ε = Greenhouse-Geisser correction

*** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.45. Summary of results of two-way repeated-measures ANOVAs with percentage correct consonant identification as the dependent variable, as a function of facewear and consonant (Experiment 4).

speech-in-noise (Experiment 4)						
within-subject factor	***F*- value / *df*s**	***p*- value**	η_p^2	***F*- value / *df*s**	***p*- value**	η_p^2
	control					
modality	$F(1,38) = 146.09$	**.000 *****	.79			
consonant	$F(9,349) = 69.88$[c]	**.000 *****	.65			
modality * consonant	$F(10,363) = 9.19$[d]	**.000 *****	.20			
	balaclava (mouth hole)			**balaclava (no mouth hole)**		
modality	$F(1,38) = 130.64$	**.000 *****	.78	$F(1,38) = 7.80$	**.008 ****	.17
consonant	$F(9,350) = 50.62$[b]	**.000 *****	.57	$F(15,570) = 47.86$	**.000 *****	.56
modality * consonant	$F(15,570) = 9.21$	**.000 *****	.20	$F(10,371) = 1.11$[a]	**.353 *ns***	.03
	motorcycle helmet			**hoodie/scarf combination**		
modality	$F(1,38) = .09$	**.762 *ns***	.00	$F(1,38) = 6.33$	**.016 ***	.14
consonant	$F(9,327) = 9.11$[e]	**.000 *****	.19	$F(10,368) = 49.55$[g]	**.000 *****	.57
modality * consonant	$F(10,364) = .52$[f]	**.869 *ns***	.01	$F(15,570) = 1.80$	**.032 ***	.05
	niqāb			**rubber mask**		
modality	$F(1,38) = .49$	**.488 *ns***	.01	$F(1,38) = .39$	**.536 *ns***	.01
consonant	$F(9,332) = 64.04$[h]	**.000 *****	.63	$F(9,344) = 40.88$[i]	**.000 *****	.52
modality * consonant	$F(15,570) = .70$	**.790 *ns***	.02	$F(10,362) = 4.78$[j]	**.000 *****	.11
	surgical mask			**tape**		
modality	$F(1,38) = 8.12$	**.007 ****	.18	$F(1,38) = 134.77$	**.000 *****	.78
consonant	$F(10,374) = 31.14$[k]	**.000 *****	.45	$F(9,326) = 63.60$[m]	**.000 *****	.63
modality * consonant	$F(10,375) = 1.93$[l]	**.041 ***	.05	$F(9,330) = 15.69$[n]	**.000 *****	.29

[a] $X^2(119) = 154.85$, $p < .05$, $\varepsilon = .65$; [b] $X^2(119) = 164.24$, $p < .01$, $\varepsilon = .61$; [c] $X^2(119) = 176.97$, $p < .01$, $\varepsilon = .61$
[d] $X^2(119) = 189.73$, $p < .001$, $\varepsilon = .64$; [e] $X^2(119) = 227.93$, $p < .001$, $\varepsilon = .57$; [f] $X^2(119) = 161.27$, $p < .01$, $\varepsilon = .64$
[g] $X^2(119) = 181.90$, $p < .001$, $\varepsilon = .65$; [h] $X^2(119) = 211.46$, $p < .001$, $\varepsilon = .58$; [i] $X^2(119) = 204.85$, $p < .001$, $\varepsilon = .60$
[j] $X^2(119) = 192.44$, $p < .001$, $\varepsilon = .64$; [k] $X^2(119) = 160.10$, $p < .05$, $\varepsilon = .66$; [l] $X^2(119) = 156.39$, $p < .05$, $\varepsilon = .66$
[m] $X^2(119) = 236.70$, $p < .001$, $\varepsilon = .57$; [n] $X^2(119) = 282.55$, $p < .001$, $\varepsilon = .58$
ε = Greenhouse-Geisser correction, *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.46. Summary of results of two-way repeated-measures ANOVAs with percentage correct consonant identification as the dependent variable, for control and each facewear condition separately, as a function of modality and consonant (Experiment 4).

quiet listening condition (Experiment 3)							
within-subject factor	***F*- value / *df*s**	***p*- value**		η_p^2	***F*- value / *df*s**	***p*- value**		η_p^2
	plosive				**fricative**			
modality	F (1,42) = .29	**.593**	***ns***	.01	F (1,42) = .20	**.656**	***ns***	.01
facewear	F (5,227) = 2.49[k]	**.028**	*	.06	F (6,240) = 56.75[g]	**.000**	***	.58
modality *facewear	F (6,232) = 2.01[l]	**.070**	***ns***	.05	F (5,224) = 1.99[h]	**.077**	***ns***	.05
	nasal				**bilabial**			
modality	F (1,42) = .00	**.954**	***ns***	.00	F (1,42) = .057	**.813**	***ns***	.00
facewear	F (5,227) = 696.47[i]	**.000**	***	.94	F (4,152) = 248.36[c]	**.000**	***	.86
modality *facewear	F (6,238) = 1.40[j]	**.221**	***ns***	.03	F (5,208) = 3.23[d]	**.008**	**	.07
	labiodental				**dental**			
modality	F (1,42) = 6.16	**.017**	*	.13	F (1,42) = 13.01	**.001**	**	.24
facewear	F (8,336) = 21.03	**.000**	***	.33	F (6,245) = 5.21[e]	**.000**	***	.11
modality *facewear	F (8,336) = 1.76	**.104**	***ns***	.04	F (6,255) = 2.60[f]	**.018**	*	.05
	alveolar				**postalveolar**			
modality	F (1,42) = .21	**.652**	***ns***	.01	F (1,42) = .25	**.620**	***ns***	.01
facewear	F (5,208) = 63.64[a]	**.000**	***	.60	F (4,182) = 56.74[m]	**.000**	***	.58
modality *facewear	F (6,234) = 2.44[b]	**.030**	*	.06	F (6,246) = .60[n]	**.729**	***ns***	.01
	velar				**voicing**			
modality	F (1,42) = .29	**.593**	***ns***	.01	F (1,42) = .31	**.579**	***ns***	.01
facewear	F (4,185) = 5.99[o]	**.000**	***	.13	F (8,336) = 5.11	**.000**	***	.11
modality *facewear	F (4,184) = .48[p]	**.764**	***ns***	.01	F (6,253) = .64[q]	**.703**	***ns***	.02

[a] X^2 (35) = 91.14, p < .001, ε = .62; [b] X^2 (35) = 83.04, p < .001, ε = .70; [c] X^2 (35) = 158.53, p < .001, ε = .45
[d] X^2 (35) = 136.89, p < .001, ε = .62; [e] X^2 (35) = 56.30, p < .05, ε = .73; [f] X^2 (35) = 50.96, p < .05, ε = .76
[g] X^2 (35) = 70.21, p < .001, ε = .71; [h] X^2 (35) = 85.14, p < .001, ε = .67; [i] X^2 (35) = 90.36, p < .001, ε = .68
[j] X^2 (35) = 76.49, p < .001, ε = .70; [k] X^2 (35) = 74.29, p < .001, ε = .68; [l] X^2 (35) = 73.39, p < .001, ε = .69
[m] X^2 (35) = 114.64, p < .001, ε = .54; [n] X^2 (35) = 60.14, p < .01, ε = .73; [o] X^2 (35) = 137.81, p < .001, ε = .55
[p] X^2 (35) = 141.36, p < .001, ε = .55; [q] X^2 (35) = 55.95, p < .05, ε = .75
ε = Greenhouse-Geisser correction, *** p < .001, ** p < .01, * p < .05, *ns* = non-significant

Table D.47. Summary of results of two-way repeated-measures ANOVAs with *d'* as the dependent variable, as a function of facewear and modality, for each phonetic feature separately (Experiment 3).

speech-in-noise (Experiment 4)						
within-subject factor	***F*- value / *df*s**	***p*- value**	η_p^2	***F*- value / *df*s**	***p*- value**	η_p^2
	plosive			**fricative**		
modality	$F(1,38) = 37.50$	**.000 *****	.50	$F(1,38) = 67.89$	**.000 *****	.64
facewear	$F(5,206) = 60.10$[d]	**.000 *****	.61	$F(7,221) = 94.19$[a]	**.000 *****	.71
modality *facewear	$F(8,304) = 5.38$	**.000 *****	.12	$F(8,304) = 32.16$	**.000 *****	.46
	nasal			**bilabial**		
modality	$F(1,38) = .59$	**.448** ***ns***	.02	$F(1,38) = 202.17$	**.000 *****	.84
facewear	$F(8,304) = 37.10$	**.000 *****	.49	$F(8,304) = 102.87$	**.000 *****	.73
modality *facewear	$F(6,218) = 1.04$[c]	**.397** ***ns***	.03	$F(8,304) = 51.22$	**.000 *****	.57
	labiodental			**dental**		
modality	$F(1,38) = 109.87$	**.000 *****	.74	$F(1,38) = 140.06$	**.000 *****	.79
facewear	$F(8,304) = 89.91$	**.000 *****	.70	$F(8,304) = 69.58$	**.000 *****	.65
modality *facewear	$F(6,218) = 37.16$[b]	**.000 *****	.49	$F(8,304) = 30.37$	**.000 *****	.44
	alveolar			**postalveolar**		
modality	$F(1,38) = 63.78$	**.000 *****	.63	$F(1,38) = 22.68$	**.000 *****	.37
facewear	$F(8,304) = 84.30$	**.000 *****	.69	$F(8,304) = 68.61$	**.000 *****	.64
modality *facewear	$F(8,304) = 11.78$	**.000 *****	.24	$F(8,304) = 5.57$	**.000 *****	.13
	velar			**voicing**		
modality	$F(1,38) = 80.51$	**.000 *****	.68	$F(1,38) = 38.59$	**.000 *****	.50
facewear	$F(8,304) = 85.69$	**.000 *****	.69	$F(7,214) = 60.77$[e]	**.000 *****	.62
modality *facewear	$F(8,304) = 4.10$	**.000 *****	.10	$F(6,220) = 4.2$[f]	**.001 ****	.10

[a] $X^2(35) = 73.86$, $p < .001$, $\varepsilon = .73$; [b] $X^2(35) = 52.59$, $p < .05$, $\varepsilon = .72$; [c] $X^2(35) = 64.89$, $p < .01$, $\varepsilon = .72$
[d] $X^2(35) = 68.05$, $p < .01$, $\varepsilon = .68$; [e] $X^2(35) = 71.99$, $p < .001$, $\varepsilon = .71$; [f] $X^2(35) = 63.52$, $p < .01$, $\varepsilon = .73$
ε = Greenhouse-Geisser correction, *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.48. Summary of results of two-way repeated-measures ANOVAs with *d'* as the dependent variable, as a function of facewear and modality, for each phonetic feature separately (Experiment 4).

response accuracy (Experiment 5)			
within-subject factor	***F*- value / *df*s**	***p*- value**	η_p^2
facewear	F (2,46) = 234.27	**.000 *** **	.91
consonant	F (5,115) = 9.54	**.000 *** **	.29
pair	F (6,137) = 35.75[a]	**.000 *** **	.61
order	F (1,23) = 9.55	**.005 ** **	.29
facewear * consonant	F (10,230) = 6.12	**.000 *** **	.21
facewear * pair	F (10,230) = 16.56[b]	**.000 *** **	.42
facewear * order	F (2,46) = 7.40	**.002 ** **	.24
consonant * pair	F (55,1265) = 4.14	**.000 *** **	.15
consonant * order	F (5,115) = 4.84	**.000 *** **	.17
pair * order	F (6,149) = 3.49[c]	**.002 ** **	.13
facewear * consonant * pair	F (110,2530) = 2.81	**.000 *** **	.11
facewear * consonant * pair * order	F (110,2530) = .99	**.504 *ns***	.04
consonant * pair * order	F (55,1265) = 1.53	**.009 ** **	.06
pair * order * facewear	F (22,506) = 2.19	**.001 ** **	.09
order * facewear * consonant	F (6,129) = 1.10[d]	**.375 *ns***	.05

[a] X^2 (65) = 97.58, $p < .01$, $\varepsilon = .54$; [b] X^2 (252) = 335.94, $p < .01$, $\varepsilon = .46$

[c] X^2 (65) = 90.93, $p < .05$, $\varepsilon = .59$; [d] X^2 (54) = 76.85, $p < .05$, $\varepsilon = .56$

ε = Greenhouse-Geisser correction; *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.49. Summary of results of a four-way repeated-measures ANOVA with response accuracy as the dependent variable, as a function of facewear, consonant, order and pair.

response time (Experiment 5)			
within-subject factor	***F*- value / *df*s**	***p*- value**	η_p^2
facewear	$F(1,31) = 32.75$[a]	**.000 *** **	.59
consonant	$F(5,115) = 1.36$	**.246 *ns***	.06
pair	$F(6,136) = 5.98$[b]	**.000 *** **	.21
order	$F(1,23) = 7.65$	**.011 ***	.25
facewear * consonant	$F(10,230) = 1.60$	**.109 *ns***	.07
facewear * pair	$F(11,242) = 4.50$[c]	**.000 *** **	.16
facewear * order	$F(2,46) = .15$	**.859 *ns***	.01
consonant * pair	$F(55,1265) = 1.43$	**.023 ***	.06
consonant * order	$F(5,115) = 1.92$	**.096 *ns***	.08
pair * order	$F(6,145) = 2.90$[d]	**.009 ** **	.11
facewear * consonant * pair	$F(110,2530) = 1.26$	**.036 ***	.05
facewear * consonant * pair * order	$F(110,2530) = 1.02$	**.431 *ns***	.04
consonant * pair * order	$F(55,1265) = 1.50$	**.012 ***	.06
pair * order * facewear	$F(22,506) = 1.66$	**.031 ***	.07
order * facewear * consonant	$F(10,230) = 1.35$	**.205 *ns***	.06

[a] $X^2(2) = 14.48$, $p < .01$, $\varepsilon = .68$; [b] $X^2(65) = 99.40$, $p < .01$, $\varepsilon = .54$

[c] $X^2(252) = 355.00$, $p < .001$, $\varepsilon = .48$; [d] $X^2(65) = 105.45$, $p < .01$, $\varepsilon = .57$

ε = Greenhouse-Geisser correction; *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.50. Summary of results of a four-way repeated-measures ANOVA with response time as the dependent variable, as a function of facewear, consonant, order and pair.

response accuracy (Experiment 5)				
within-subject factor	***F*- value / *df* s**	***p*- value**		η_p^2
		control		
consonant	F (5,115) = 3.10	**.012**	*	.12
pair	F (11,253) = 4.02	**.000**	***	.15
order	F (1,23) = .29	**.595**	***ns***	.01
consonant * pair	F (55,1265) = 1.91	**.000**	***	.08
consonant * order	F (5,115) = 1.40	**.231**	***ns***	.06
pair * order	F (6,136) = 2.01[a]	**.069**	***ns***	.08
consonant * pair * order	F (55,1265) = 1.04	**.390**	***ns***	.04
		helmet		
consonant	F (5,115) = 7.55	**.000**	***	.25
pair	F (11,253) = 15.79	**.000**	***	.41
order	F (1,23) = 9.78	**.005**	**	.30
consonant * pair	F (55,1265) = 3.22	**.000**	***	.12
consonant * order	F (5,115) = 3.94	**.002**	**	.15
pair * order	F (11,253) = 3.14	**.001**	**	.12
consonant * pair * order	F (55,1265) = .97	**.539**	***ns***	.04
		tape		
consonant	F (5,115) = 8.23	**.000**	***	.26
pair	F (11,253) = 39.65	**.000**	***	.63
order	F (1,23) = 9.96	**.004**	**	.30
consonant * pair	F (55,1265) = 3.81	**.000**	***	.14
consonant * order	F (3,80) = 1.86[b]	**.135**	***ns***	.08
pair * order	F (11,253) = 2.57	**.004**	**	.10
consonant * pair * order	F (55,1265) = 1.39	**.033**	*	.06

[a] X^2 (65) = 97.82, $p < .01$, $\varepsilon = .54$; [b] X^2 (14) = 24.72, $p < .05$, $\varepsilon = .70$

ε = Greenhouse-Geisser; *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.51. Summary of results of three-way repeated-measures ANOVAs with response accuracy as the dependent variable, for control, helmet, and tape separately, as a function of consonant, pair, and order.

response time (Experiment 5)				
within-subject factor	***F*- value / *df* s**	***p*- value**		η_p^2
		control		
consonant	$F(5,115) = 1.50$	**.196**	***ns***	.06
pair	$F(6,149) = 1.42$[a]	**.207**	***ns***	.06
order	$F(1,23) = 4.86$	**.038**	*	.17
consonant * pair	$F(55,1265) = 1.57$	**.005**	**	.06
consonant * order	$F(5,115) = .71$	**.617**	***ns***	.03
pair * order	$F(11,253) = 1.31$	**.219**	***ns***	.05
consonant * pair * order	$F(55,1265) = 1.09$	**.309**	***ns***	.05
		helmet		
consonant	$F(5,115) = .88$	**.497**	***ns***	.04
pair	$F(6,149) = 2.56$[b]	**.019**	*	.10
order	$F(1,23) = 5.54$	**.027**	*	.19
consonant * pair	$F(55,1265) = 1.12$	**.254**	***ns***	.05
consonant * order	$F(3,75) = 3.17$[c]	**.026**	*	.12
pair * order	$F(6,139) = 1.53$[d]	**.172**	***ns***	.06
consonant * pair * order	$F(55,1265) = .92$	**.640**	***ns***	.04
		tape		
consonant	$F(3,79) = 2.14$[e]	**.093**	***ns***	.09
pair	$F(5,126) = 9.95$[f]	**.000**	***	.30
order	$F(1,23) = 3.39$	**.079**	***ns***	.13
consonant * pair	$F(55,1265) = 1.32$	**.063**	***ns***	.05
consonant * order	$F(5,115) = .96$	**.447**	***ns***	.04
pair * order	$F(11,253) = 3.26$	**.000**	***	.12
consonant * pair * order	$F(55,1265) = 1.49$	**.013**	*	.06

[a] $X^2(65) = 107.98$, $p < .001$, $\varepsilon = .59$; [b] $X^2(65) = 112.40$, $p < .001$, $\varepsilon = .59$

[c] $X^2(14) = 30.59$, $p < .01$, $\varepsilon = .66$; [d] $X^2(65) = 104.46$, $p < .01$, $\varepsilon = .55$

[e] $X^2(14) = 24.75$, $p < .05$, $\varepsilon = .69$; [f] $X^2(65) = 100.91$, $p < .01$, $\varepsilon = .50$

ε = Greenhouse-Geisser; *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* = non-significant

Table D.52. Summary of results of three-way repeated-measures ANOVAs with response time as the dependent variable, for control, helmet, and tape separately, as a function of consonant, pair, and order.

D.6 *T*-tests

participant	% correct	std. error mean	*t*- value
01	82.4	2.0	16.546 ***
02	78.2	2.2	12.615 ***
03	81.3	2.1	15.190 ***
04	81.0	2.1	14.866 ***
05	78.7	2.4	12.201 ***
06	82.4	1.9	17.031 ***
07	81.7	2.1	15.466 ***
08	78.5	2.1	13.473 ***
09	78.5	2.2	13.160 ***
10	74.5	2.2	11.405 ***
11	71.5	2.3	9.423 ***
12	74.3	2.3	10.584 ***
13	81.9	2.0	15.606 ***
14	77.8	2.2	12.527 ***
15	76.9	2.3	11.830 ***
16	79.4	2.1	13.964 ***
17	72.7	2.2	10.109 ***
18	71.1	2.2	9.535 ***
19	64.6	2.3	6.461 ***
20	81.9	2.2	14.861 ***
21	69.9	2.2	9.056 ***
22	80.3	1.9	16.143 ***
23	88.9	1.7	23.394 ***
24	87.3	1.8	21.276 ***

*** $p < .001$, all dfs = 215

Table D.53. Response accuracy (percentage correct/standard error of the mean) averaged across facewear, consonant, pair, and order, for each of the 24 participants in Experiment 5 separately. The rightmost column shows *t*-values derived from a series of one-sample *t*-tests. The *p*-values (all $p < .001$) indicate that the mean talker discrimination accuracy obtained by all participants was significantly higher than chance level (50%).

consonant	facewear	% correct	std. error mean	*t*- value
	control	96.0	1.3	35.0 ***
/t/	helmet	78.1	2.2	13.0 ***
	tape	66.2	1.7	9.6 ***
	control	91.5	1.9	22.2 ***
/p/	helmet	70.3	2.3	8.8 ***
	tape	59.2	1.8	5.0 ***
	control	92.5	1.7	25.6 ***
/s/	helmet	76.7	2.2	12.0 ***
	tape	71.0	1.7	12.4 ***
	control	91.8	1.6	25.9 ***
/f/	helmet	78.3	1.8	15.8 ***
	tape	66.2	2.1	7.6 ***
	control	91.7	1.6	25.6 ***
/n/	helmet	74.3	2.2	10.9 ***
	tape	73.3	2.3	10.0 ***
	control	92.2	1.5	27.9 ***
/m/	helmet	67.7	1.7	10.3 ***
	tape	69.8	1.8	10.7 ***

*** $p < .001$, all dfs = 23

Table D.54. Response accuracy (percentage correct/standard error of the mean) averaged across participants, as a function of facewear, for each of the six consonants tested in Experiment 5 separately. The rightmost column shows *t*-values derived from a series of one-sample *t*-tests. The *p*-values (all $p < .001$) indicate that the mean talker discrimination accuracy was for all consonants in all facewear conditions significantly higher than chance level (50%).

D.7 Illustrations

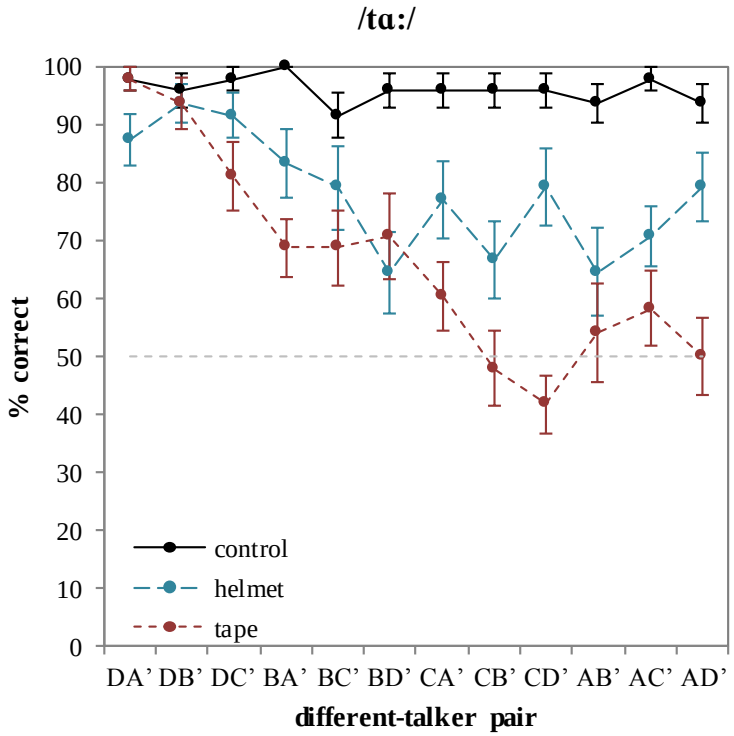

Figure D.2. Response accuracy for all twelve different-talker pairs as a function of facewear, for the test syllable /tɑ:/ separately. The dashed horizontal line represents chance level (50%). The error bars show the standard error of the mean.

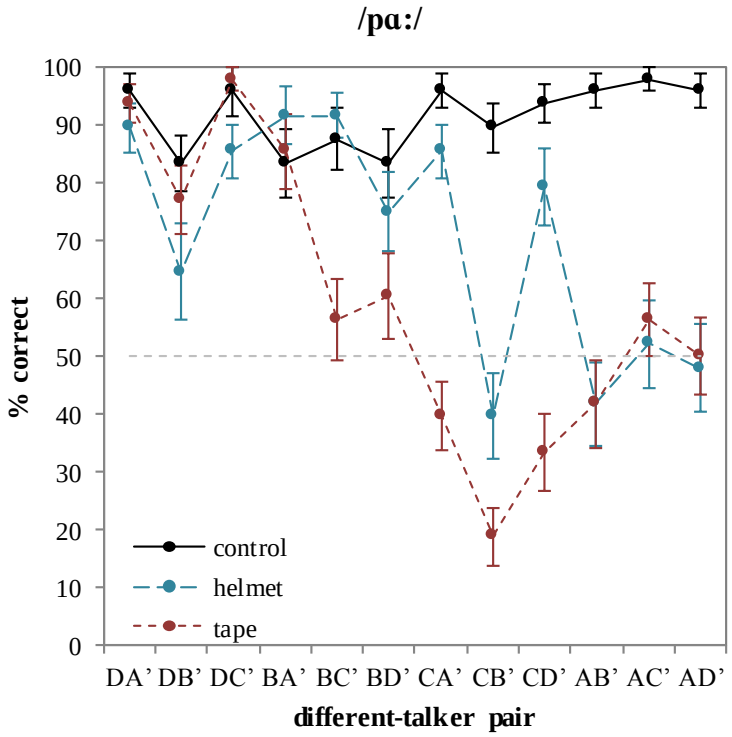

Figure D.3. Response accuracy for all twelve different-talker pairs as a function of facewear, for the test syllable /pɑː/ separately. The dashed horizontal line represents chance level (50%). The error bars show the standard error of the mean.

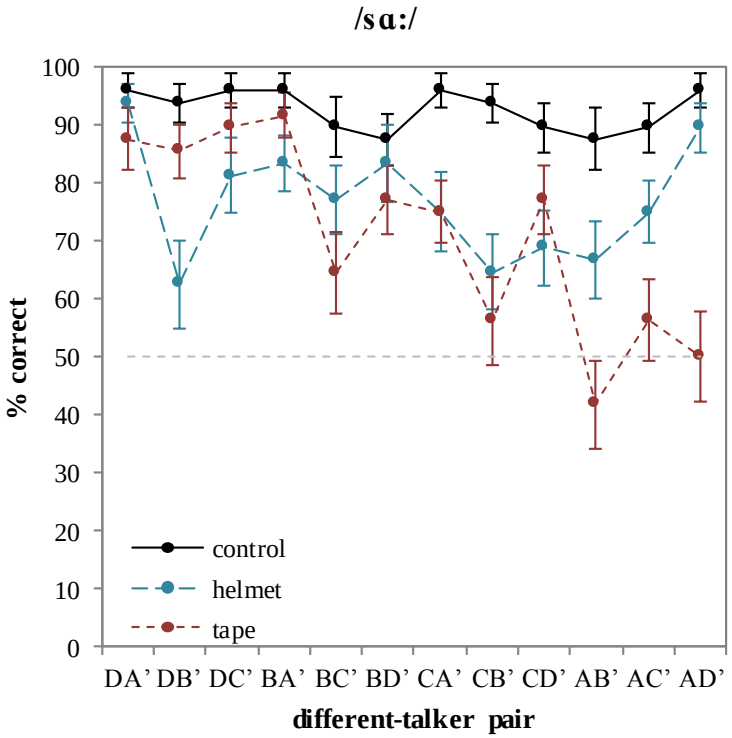

Figure D.4. Response accuracy for all twelve different-talker pairs as a function of facewear, for the test syllable /sɑː/ separately. The dashed horizontal line represents chance level (50%). The error bars show the standard error of the mean.

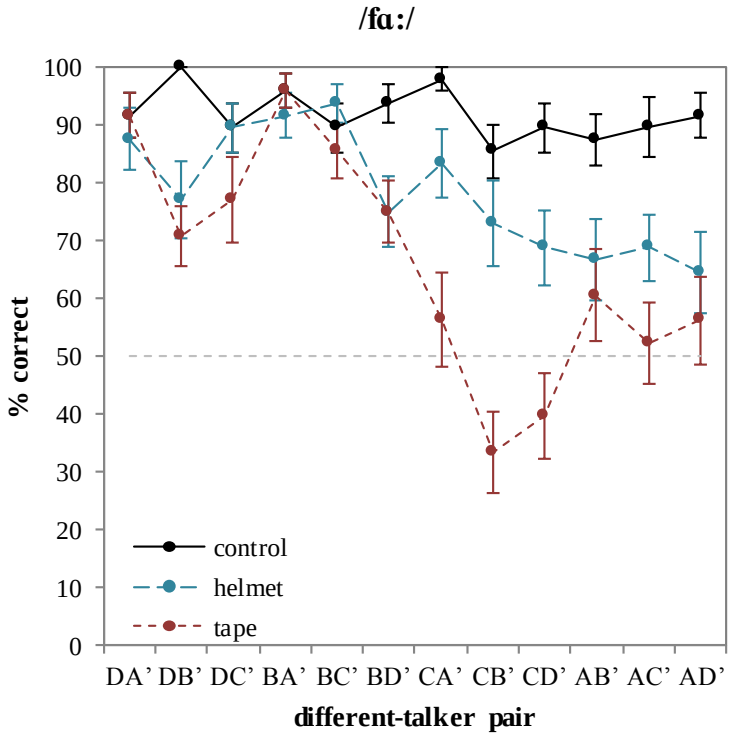

Figure D.5. Response accuracy for all twelve different-talker pairs as a function of facewear, for the test syllable /fɑ:/ separately. The dashed horizontal line represents chance level (50%). The error bars show the standard error of the mean.

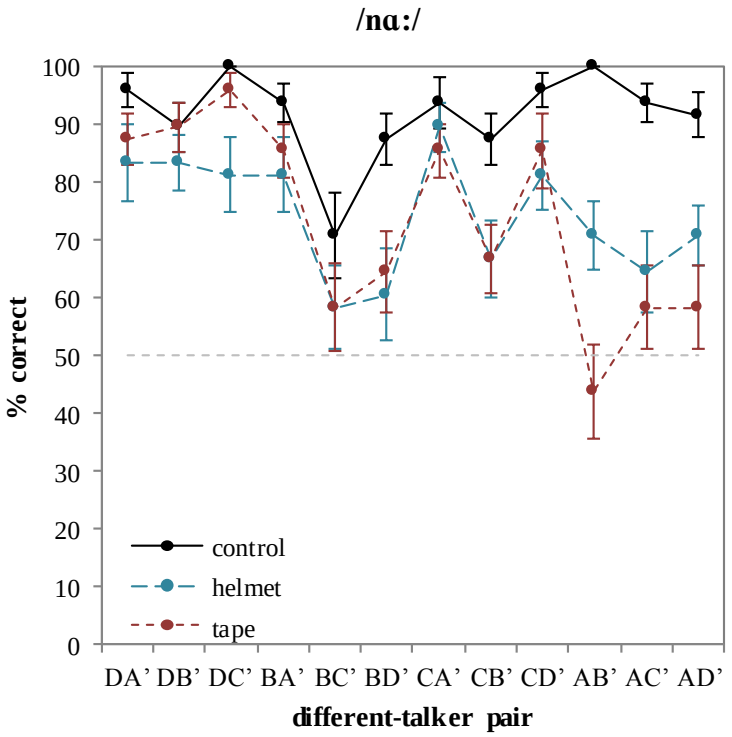

Figure D.6. Response accuracy for all twelve different-talker pairs as a function of facewear, for the test syllable /nɑ:/ separately. The dashed horizontal line represents chance level (50%). The error bars show the standard error of the mean.

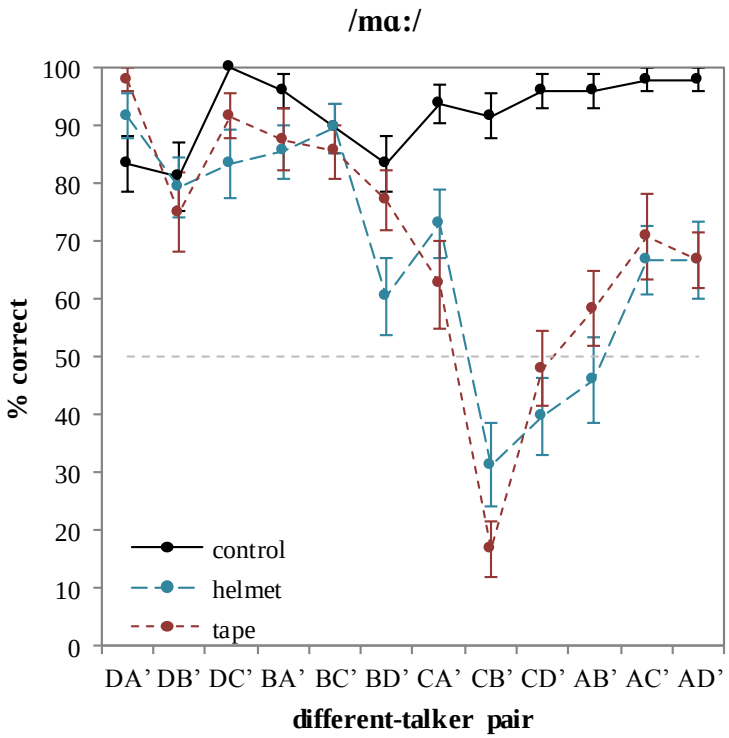

Figure D.7. Response accuracy for all twelve different-talker pairs as a function of facewear, for the test syllable /mɑː/ separately. The dashed horizontal line represents chance level (50%). The error bars show the standard error of the mean.

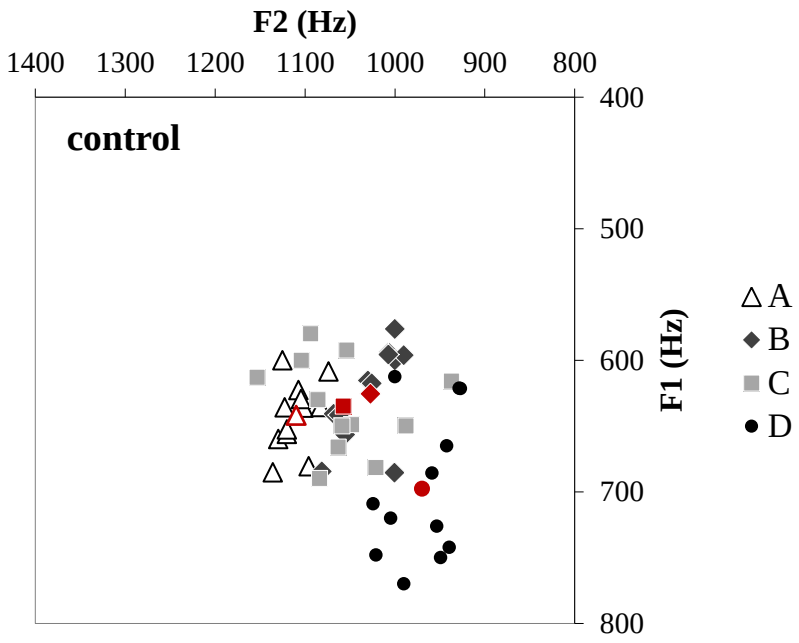

Figure D.8. Mean first formants (F1) and second formants (F2) of /ɑː/ (in Hz) produced by talkers A, B, C, and D in the control condition. The red data points indicate the means of each talker's F1 and F2 values (the means are also shown in Figure 6.6 in Chapter 6).

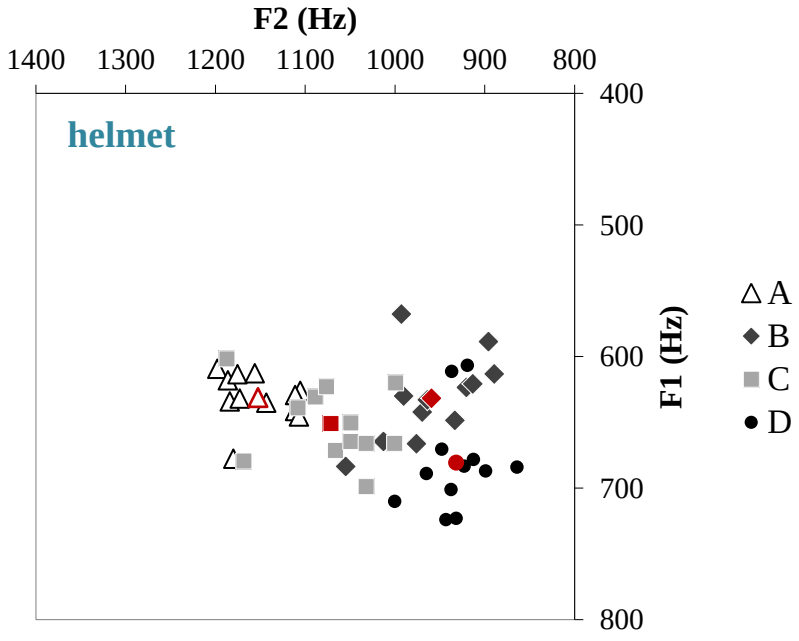

Figure D.9. Mean first formants (F1) and second formants (F2) of /ɑː/ (in Hz) produced by talkers A, B, C, and D in the helmet condition. The red data points indicate the means of each talker's F1 and F2 values (the means are also shown in Figure 6.6 in Chapter 6).

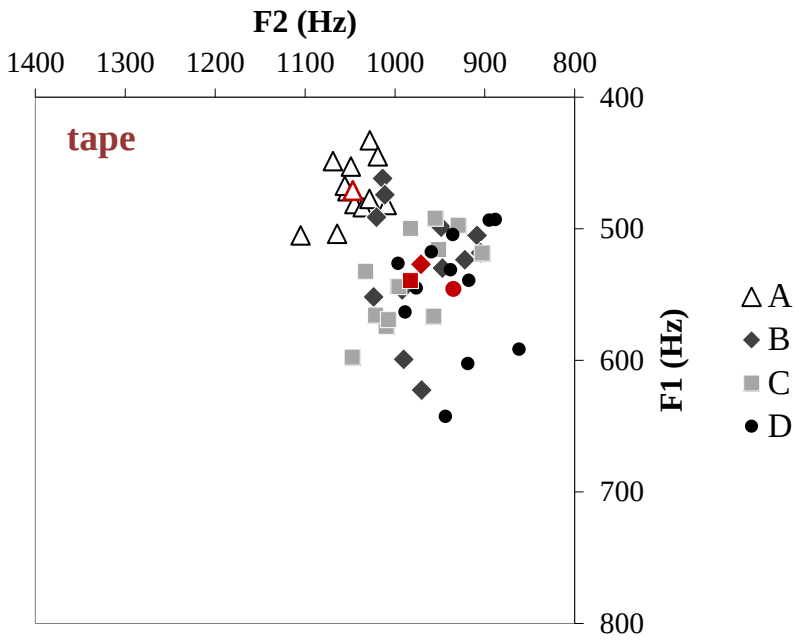

Figure D.10. Mean first formants (F1) and second formants (F2) of /ɑː/ (in Hz) produced by talkers A, B, C, and D in the tape condition. The red data points indicate the means of each talker's F1 and F2 values (the means are also shown in Figure 6.6 in Chapter 6).

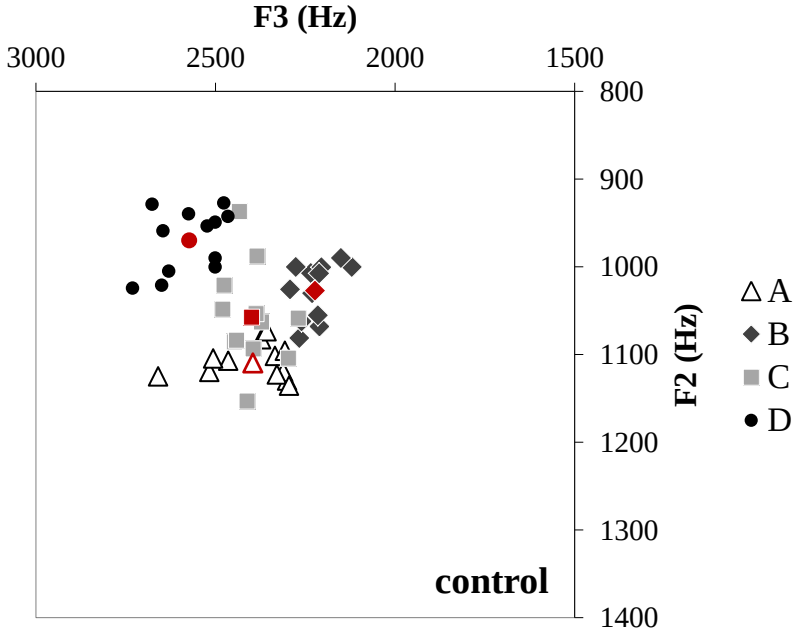

Figure D.11. Mean second formants (F2) and third formants (F3) of /ɑː/ (in Hz) produced by talkers A, B, C, and D in the control condition. The red data points indicate the means of each talker's F2 and F3 values (the means are also shown in Figure 6.7 in Chapter 6).

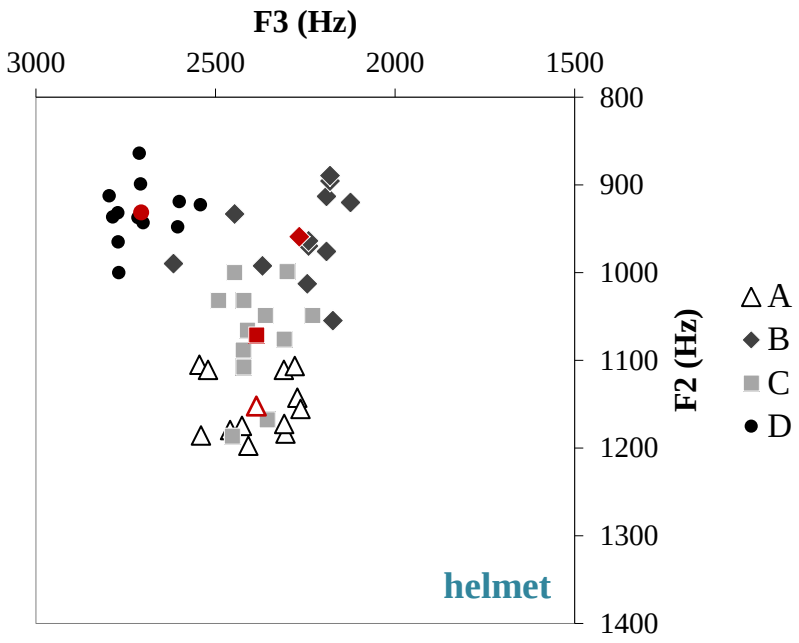

Figure D.12. Mean second formants (F2) and third formants (F3) of /ɑː/ (in Hz) produced by talkers A, B, C, and D in the helmet condition. The red data points indicate the means of each talker's F2 and F3 values (the means are also shown in Figure 6.7 in Chapter 6).

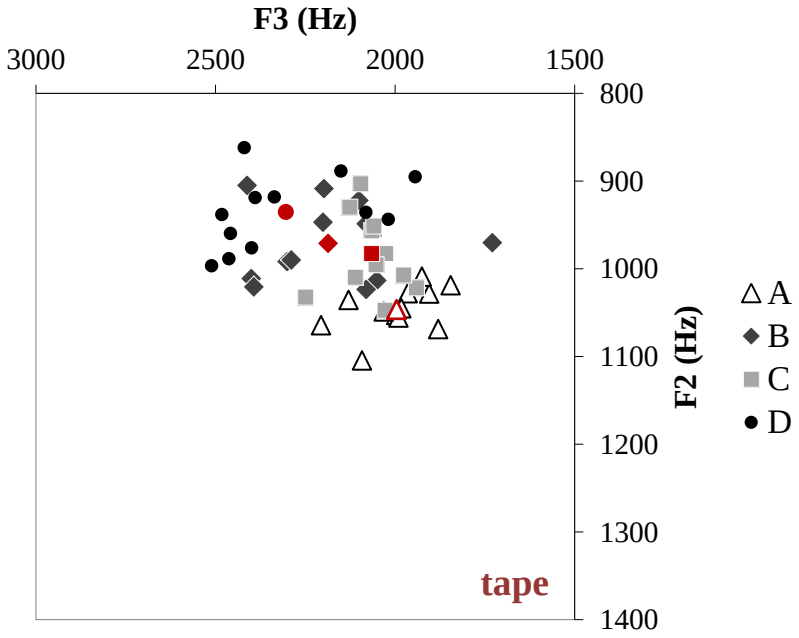

Figure D.13. Mean second formants (F2) and third formants (F3) of /ɑː/ (in Hz) produced by talkers A, B, C, and D in the tape condition. The red data points indicate the means of each talker's F2 and F3 values (the means are also shown in Figure 6.7 in Chapter 6).

D.8 Correlations

consonant	Pearson's *r*	*p* -value
control x helmet		
/t/	.306	.333 *ns*
/p/	–.234	.465 *ns*
/s/	.374	.231 *ns*
/f/	.276	.386 *ns*
/n/	.599	.040 *
/m/	–.219	.493 *ns*
control x tape		
/t/	.318	.314 *ns*
/p/	–.154	.632 *ns*
/s/	.447	.146 *ns*
/f/	.394	.205 *ns*
/n/	.289	.362 *ns*
/m/	–.187	.560 *ns*
helmet x tape		
/t/	.655	.021 *
/p/	.568	.054 *ns*
/s/	.277	.384 *ns*
/f/	.760	.004 **
/n/	.789	.002 **
/m/	.909	.000 ***

*** $p < .001$, ** $p < .01$,* $p < .05$, *ns* = non-significant, all $N = 12$

Table D.55. Pearson correlation coefficients computed between the mean accuracy score per pair in the control and each of the facewear conditions, as well as between the scores for the two facewear conditions, for each consonant separately (Experiment 5).

References

Abel, S. M., Alberti, P. W. & Riko, K. (1980). Speech intelligibility in noise with ear protectors. *Journal of Otolaryngology* **9**(3), 256–265.

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Abeysekera, J. D. & Shahnavaz, H. (1987). Ergonomics assessment of selected dust respirators: Their use in the tropics. *Applied Ergonomics* **18**(4), 266–272.

Aleksic, P. S. & Katsaggelos, A. G. (2006). Audio-visual biometrics. *Proceedings of the IEEE* **94**(1), 2025–44.

Alexander, A., Dessimoz, D., Botti, F. & Drygajlo A. (2005). Aural and automatic forensic speaker recognition in mismatched conditions. *The International Journal of Speech, Language and the Law* **12**(2), 214–234.

Amino, K. & Arai, T. (2009). Speaker-dependent characteristics of the nasals. *Forensic Science International* **185**(1–3), 21–28.

Andics, A. (2013). *Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning*. Ph.D. dissertation, Radboud University Nijmegen.

Andics, A., McQueen, J. M. & Turennout, M. van (2007). Phonetic content influences voice discriminability. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, August 6–10, 2007, 1829–32.

Armitage, R. (2002). *To CCTV or not to CCTV? A review of current research into the effectiveness of CCTV systems in reducing crime* (May 2002). London: Nacro, Crime and Social Policy Section. Available from: http://www.goo.gl/UPw91n. [Accessed: 7th May 2014].

Aso, S. & Kinoshita, R. (1963a). Absorption of sound wave by fabrics, part I: Absorption mechanisms. *Journal of the Textile Machinery Society of Japan* (English edition) **9**, 1–15.

Aso, S. & Kinoshita, R. (1963b). Absorption of sound wave by fabrics, part II: Acoustic impeadence [sic] density. *Journal of the Textile Machinery Society of Japan* (English edition) **9**, 40–46.

Aso, S. & Kinoshita, R. (1964). Absorption of sound wave by fabrics, part III: Flow resistance. *Journal of the Textile Machinery Society of Japan* (English edition) **10**(5), 236–241.

Badin, P., Tarabalka, Y., Elisei, F. & Bailly, G. (2010). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech communication* **52**(6), 493–503.

Ball, K., Haggerty, K. & Lyon, D. (eds.) (2012). *Routledge Handbook of Surveillance Studies*. Abingdon, New York: Routledge.

Barker, J. & Shao, X. (2009). Energetic and informational masking effects in an audio-visual speech recognition system. *IEEE Transactions on Audio, Speech and Language Processing* **17**(3), 446–458.

Baum, S. R., McFarland, D. H. & Diab, M. (1996). Compensation to articulatory perturbation: Perceptual data. *The Journal of the Acoustical Society of America* **99**(6), 3791–4.

BBC News (2006). Lawyers 'can wear veils in court'. [Online]. 10th November 2006. Available from: http://www.goo.gl/guSMmh. [Accessed: 7th May 2014].

BBC News (2007a). School head explains niqab ban. [Online]. 12th February 2007. Available from: http://www.goo.gl/XN4RyI. [Accessed: 7th May 2014].

BBC News (2007b). Schoolgirl loses veil legal case. [Online]. 21st February 2007. Available from: http://www.goo.gl/222121. [Accessed: 7th May 2014].

BBC News (2013a). Viewpoints: Should full-face veils be banned in some public places? [Online]. 16th September 2013. Available from: http://www.goo.gl/5URbta. [Accessed: 7th May 2014].

BBC News (2013b). Selfridges robbery: 'Men in burkas' in 'smash and grab'. [Online]. 7th June 2013. Available from: http://www.goo.gl/SOa2VX. [Accessed: 7th May 2014].

BBC News (2013c). Head teachers against face veils in school. [Online]. 16th September 2013. Available from: http://www.goo.gl/nPhiwz. [Accessed: 7th May 2014].

BBC News (2014). Burka escape terror suspect begins appeal. [Online]. 27th January 2014. Available from: http://www.goo.gl/b8yaQr. [Accessed: 7th May 2014].

Benkí, J. R. (2003). Analysis of English nonsense syllable recognition in noise. *Phonetica* **60**(2), 129–157.

Benoît, C., Guiard-Marigny, T., Le Goff, B. & Adjoudani, A. (1996). Which components of the face do humans and machines best speechread? In: Stork, D. G. & Hennecke, M. E. (eds.). *Speechreading by Humans and Machines: Models, Systems, and Applications*. Berlin, Heidelberg, New York: Springer, 315–328.

Bensel, C. K., Teixeira, R. A. & Kaplan, D. B. (1987). *The effects of US army chemical protective clothing on speech intelligibility, visual field, body mobility and psychomotor coordination of men* (Technical Report Natick/TR-87/037, September 1987). Natick, Mass., USA: United States Army Natick Research, Development and Engineering Center, Individual Protection Directorate. Available from: http://www.goo.gl/UPw91n. [Accessed: 7th May 2014].

Berger, K. W., Garner, M. & Sudman, J. (1971). The effect of degree of facial exposure and the vertical angle of vision on speechreading performance. *The Teacher of the Deaf* **69**, 322–326.

Bernstein, L. E. & Auer, E. T. Jr. (1996). Word recognition in speechreading. In: Stork, D. G., Hennecke, M. E. (eds.). *Speechreading by Humans and Machines: Models, Systems, and Applications*. Berlin, Heidelberg, New York: Springer, 17–26.

Bernstein, L. E., Demorest, M. E. & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics* **62**(2), 233–252.

Betancourt, K. S. & Bahr, R. H. (2010). The influence of signal complexity on speaker identification. *The International Journal of Speech, Language and the Law* **17**(2), 179–200.

Bishop, J., Bahr, R. H. & Gelfer, M. (1989). Near-field speech intelligibility in chemical-biological warfare masks. *Military Medicine* **164**(8), 543–550.

Blacklock, O. S. & Shadle, C. H. (2003). Spectral moments and alternative methods of characterizing fricatives. *The Journal of the Acoustical Society of America* **113**(4), 2199–2199.

Blatchford, H. & Foulkes, P. (2006). Identification of voices in shouting. *The International Journal of Speech, Language and the Law* **13**(2), 241–254.

Bond, Z. S., Moore, T. J. & Gable, B. (1989). Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask. *The Journal of the Acoustical Society of America* **85**(2), 907–912.

Bricker, P. D. & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America* **40**(6), 1441–9.

Brungart, D. S. & Simpson, B. D. (2005). Interference from audio distracters during speechreading. *The Journal of the Acoustical Society of America* **118**(6), 3889–902.

Brunner, J. (2009). *Perturbed Speech. How Compensation Mechanisms can Inform us about Phonemic Targets*. Saarbrücken: Südwestdeutscher Verlag für Hochschulschriften.

Bull, R. & Clifford, B. R. (1984). Earwitness voice recognition accuracy. In: Wells, G. L. & Loftus, E. F. (eds.). *Eyewitness Testimony: Psychological Perspectives*. Cambridge: Cambridge University Press, 92–123.

Burnham, D. & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology* **45**(4), 204–220.

Burnham, D., Lau, S., Tam, H. & Schoknecht, C. (2001). Visual discrimination of Cantonese tones by tonal but non-Cantonese speakers and by non-tonal language speakers. *Proceedings of the 4th International Conference on Auditory-Visual Speech Processing (AVSP)*, Aalborg, Denmark, September 7–9, 2001, 155–160.

Byrne, C., & Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. *The International Journal of Speech, Language and the Law* **11**(1), 83–102.

Campbell, M. (2009). Innovation: The sinister powers of crowdsourcing. *New Scientist*. [Online]. 22nd December 2009. Available from: http://www.goo.gl/SBtfv3. [Accessed: 7th May 2014].

Caretti, D. M. & Strickler L. C. (2003). Speech intelligibility during respirator wear: Influences of respirator speech diaphragm size and background noise. *American Industrial Hygiene Association Journal* **64**(6), 846–850.

Casciani, D. (2013). Analysis: The niqab judgement explained. *BBC News*. [Online]. 16th September 2013. Available from: http://www.goo.gl/3zq5fu. [Accessed: 7th May 2014].

Chen, Y. & Hazan, V. (2009). Developmental factor and the non-native speaker effect in auditory-visual speech perception. *The Journal of the Acoustical Society of America* **126**(2), 858–865.

Cho, T. & Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics* **27**(2), 207–229.

Cho, T. & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics* **33**(2), 121–157.

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

Clark, J. & Foulkes, P. (2007). Identification of voices in electronically disguised speech. *The International Journal of Speech, Language and the Law* **14**(2), 195–221.

Clark, J., Yallop, C. & Fletcher, J. (2007). *An Introduction to Phonetics and Phonology*. 3rd ed. Malden, Oxford, Carlton: Blackwell Publishing.

Colantoni, L. & Marinescu, I. (2008). The scope of stop weakening in Argentine Spanish. *Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology*, Austin, TX, USA, September 26–28, 2008, 100–104.

Coniam, D. (2005). The impact of wearing a face mask in a high-stakes oral examination: An exploratory post-SARS study in Hong Kong. *Language Assessment Quarterly* **2**, 235–261.

Cook, S. & Wilding, J. (1997). Earwitness testimony 2: Voices, faces and context. *Applied Cognitive Psychology* **11**(6), 527–541.

Cook, S. & Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the face overshadowing effect. *British Journal of Psychology* **92**(4), 617–629.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America* **119**(3), 1562–73.

Cooke, M., Lecumberri, M. L. G. & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America* **123**(1), 414–427.

Cotton, J. C. (1935). Normal 'visual hearing'. *Science* **82**, 592–593.

Coyne, K. M. & Barker, D. J. (2010). *Speech intelligibility while wearing civilian full-facepiece air-purifying respirators* (Final Report ECBC-TR-779, June 2010). Aberdeen, MD, USA: U.S. Army Research, Development and Engineering Command, Chemical Biological Center, Research and Technology Directorate. Available from: http://www.goo.gl/aDafFz. [Accessed: 7th May 2014].

Coyne, K. M., Johnson, A. T., Yeni-Komshian, G. H. & Dooly, C. R. (1998). Respirator performance ratings for speech intelligibility. *American Industrial Hygiene Association Journal* **59**(4), 257–260.

Crowder, R. G. (1982). A common basis for auditory sensory storage in perception and immediate memory. *Perception & Psychophysics* **31**(5), 477–483.

Cutler, A., Andics, A. & Fang, Z. (2011). Inter-dependent categorization of voices and segments. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)*, Hong Kong, China, August 17–21, 2011, 552–555.

Cutler, A., Cooke, M., Lecumberri, M. L. G. & Pasveer, D. (2007). L2 consonant identification in noise: Cross-language comparisons. *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech)*, Antwerp, Belgium, August 27–31, 2007, 1585–8.

Cutler, A., Lecumberri, M. L. G. & Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: Effects of local context. *The Journal of the Acoustical Society of America* **124**(2), 1264–8.

Cutler, A., Weber, A., Smits, R. & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America* **116**(6), 3668–78.

Cvejic, E., Kim, J. & Davis, C. (2010). Modification of prosodic cues when an interlocutor cannot be seen: The effect of visual feedback on acoustic prosody production. *Proceedings of the 20th International Congress on Acoustics (ICA)*, Sydney, Australia, August 23–27, 2010.

Cvejic, E., Kim, J. & Davis, C. (2011). Perceiving visual prosody from point-light displays. *Proceedings of the 11th International Conference on Auditory-Visual Speech Processing (AVSP)*, Volterra, Italy, August 31–September 3, 2011, 15–20.

Davis, C. & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition* **100**(3), B21–31.

Davis, C. & Kim, J. (2009). Recognizing spoken vowels in multi-talker babble: Spectral and visual speech cues. *Proceedings of the 9th International Conference on Auditory-Visual Speech Processing (AVSP)*, Norwich, United Kingdom, September 10–13, 2009, 130–133.

Davis, C. & Kim, J. (2010). Transfer of talker-familiarity effects. *Proceedings of the 20th International Congress on Acoustics (ICA)*, Sydney, Australia, August 23–27, 2010.

Delattre, P. C., Liberman, A. M. & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America* **27**(4), 769–773.

Dellwo, V., Huckvale, M. & Ashby, M. (2007). How is individuality expressed in voice? An introduction to speech production and description for speaker classification. In: Müller, C. (ed.). *Speaker Classification I: Fundamentals, Features, and Methods*. Berlin, Heidelberg: Springer, 1–20.

Deterding, D. (2006). The North Wind versus a Wolf: Short texts for the description and measurement of English pronunciation. *Journal of the International Phonetic Association* **36**(2), 187–196.

Deterding, D. & Nolan, F. (2007). Aspiration and voicing in Chinese and English plosives. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, August 6–10, 2007, 385–388.

Docherty, G. J. (1992). *The Timing of Voicing in British English Obstruents*. Berlin, New York: Foris Publications.

Dodson, C. S., Johnson, M. K. & Schooler, J. W. (1997). The verbal overshadowing effect: Why descriptions impair face recognition. *Memory & Cognition* **25**(2), 129–139.

Dohen, M., Lœvenbruck, H., Cathiard, M.-A. & Schwartz, J.-L. (2004a). Identification of the possible visible correlates of contrastive focus in French. *Proceedings of the 2nd International Conference on Speech Prosody*, Nara, Japan, March 23–26, 2004, 73–76.

Dohen, M., Lœvenbruck, H., Cathiard, M.-A. & Schwartz, J.-L. (2004b). Visual perception of contrastive focus in reiterant French speech. *Speech Communication* **44**(1–4), 15–172.

Drager, K. (2009). Language, stance, and identity at Selwyn Girls' High. *Proceedings of the 5th Biennial International Gender and Language Association Conference (IGALA)*, Wellington, New Zealand, July 3–5, 2008, 419–433.

Drager, K., Schutz, R., Chik, I., Hardeman, K. & Jih, V. (2012). When hearing is believing: Perceptions of speaker style, gender, ethnicity, and pitch. *Paper presented at the 86th Annual Meeting of the Linguistic Society of America (LSA)*, Portland, OR, USA, January 5–8, 2012.

Dubno, J. R. & Levitt, H. (1981). Predicting consonant confusions from acoustic analysis. *The Journal of the Acoustical Society of America* **69**(1), 249–261.

Eck, E. K. & Vannier, A. (1997). The effect of high-efficiency particulate air respirator design on occupational health: A pilot study balancing risks in the real world. *Infection Control and Hospital Epidemiology* **18**(2), 122–127.

Eckert P. (1996). Vowels and nail polish: The emergence of linguistic style in the preadolescent heterosexual marketplace. *Gender and Belief Systems: Proceedings of the 4th Berkeley Women and Language Conference*, Berkeley, CA, USA, April 19–21, 1996, 183–190.

Erber, N. P. (1974). Effects of angle, distance, and illumination on visual reception of speech by profoundly deaf children. *Journal of Speech, Language, and Hearing Research* **17**(1), 99–112.

Erber, N. P. (1979). Auditory-visual perception of speech with reduced optical clarity. *Journal of Speech, Language, and Hearing Research* **22**(2), 212–223.

Eriksson, E. J., Sullivan, K. P. H., Zetterholm, E., Czigler, P. E., Green, J., Skagerstrand, Å. & Doorn, J. van (2010). Detection of imitated voices: Who are reliable earwitnesses? *The International Journal of Speech, Language and the Law* **171**(1), 25–44.

Fagel, S. (2005). Auditory speech illusion evoked by moving lips. *Proceedings of the 10th International Conference on Speech and Computer (SPECOM)*, Patras, Greece, October 17–19, 2005, 115–118.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton Publishers.

Fecher, N. (2008). *Der Einfluss von Telefonsprache auf die Akustik von Vokalen*. Magisterarbeit, Ludwig-Maximilians-Universität München.

Fecher, N. (2011). Spectral properties of fricatives: a forensic approach. *Proceedings of the 4th ISCA Tutorial and Research Workshop on Experimental Linguistics (ExLing)*, Paris, France, May 25–27, 2011, 71–74.

Fecher, N. (2012). The 'Audio-Visual Face Cover Corpus': Investigations into audio-visual speech and speaker recognition when the speaker's face is occluded by facewear. *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech)*, Portland, OR, USA, September 9–13, 2012.

Fecher, N. & Watt, D. (2011). Speaking under cover: The effect of face-concealing garments on spectral properties of fricatives. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)*, Hong Kong, China, August 17–21, 2011, 663–666.

Fecher, N. & Watt, D. (2013). Effects of forensically-realistic facial concealment on auditory-visual consonant recognition in quiet and noise conditions. *Proceedings of the 12th International Conference on Auditory-Visual Speech Processing (AVSP)*, Annecy, France, August 29 – September 1, 2013.

Fellowes, J. M., Remez, R. E. & Rubin, P. E. (1996). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics* **59**(6), 839–849.

Firth, N. (2011). Face recognition technology fails to find UK rioters. *New Scientist*. [Online]. 18 August 2011. Available from: http://www.goo.gl/kyyx0z. [Accessed: 7th May 2014].

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech, Language, and Hearing Research* **11**(4), 796–804.

Fitzpatrick, M. & Kim, J. (2010). Audio-visual speech perception in noise by first and second language listeners. *Proceedings of 20th International Congress on Acoustics (ICA)*, Sydney, Australia, August 23–27, 2010.

Fleming, N. (2011). Smartphone surveillance: The cop in your pocket. *New Scientist*. [Online]. 3rd August 2011. Available from: http://www.goo.gl/wPtpr7. [Accessed: 7th May 2014].

Flipsen, P. Jr., Shriberg, L., Weismer, G., Karlsson, H. & McSweeny, J. (1999). Acoustic characteristics of /s/ in adolescents. *Journal of Speech, Language, and Hearing Research* **42**(3), 663–677.

Forrest, K., Weismer, G., Milenkovic, P. & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America* **84**(1), 115–123.

Foulkes, P. & Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *The International Journal of Speech, Language and the Law* **7**(2), 180–198.

Foulkes, P., Docherty, G. & Jones, M. J. (2010). Analyzing stops. In: Di Paolo, M. & Yaeger-Dror, M. (eds.). *Sociophonetics: A Student's Guide*. London: Routledge, 58–71.

Foulkes, P. & French, P. (2012). Forensic speaker comparison: A linguistic-acoustic perspective. In: Tiersma, P. M. & Solan, L. M. (eds.). *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press, 557–572.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* **14**, 3–28.

Fraser, H. (2003). Issues in transcription: Factors affecting the reliability of transcripts as evidence in legal cases. *The International Journal of Speech, Language and the Law* **10**(2), 203–226.

Fraser, H. & Stevenson, B. (2014). The power and persistence of contextual priming: More risks in using police transcripts to aid jurors' perception of poor quality covert recordings. *The International Journal of Evidence & Proof* **18**(3), 205–229.

Fraser, H., Stevenson, B. & Marks, T. (2011). Interpretation of a crisis call: Persistence of a primed perception of a disputed utterance. *The International Journal of Speech, Language and the Law* **18**(2), 261–292.

French, J. P. (1990). Analytic procedures for the determination of disputed utterances. In: Kniffka, H. (ed.). *Texte zu Theorie und Praxis forensischer Linguistik*. Tübingen: Max Niemeyer Verlag, 201–213.

French, J. P. & Harrison, P. (2006). Investigative and evidential application of forensic speech science. In: Heaton-Armstrong, A., Shepherd, E., Gudjonsson, G. & Wolchover, D. (eds.). *Witness Testimony: Psychological, Investigative and Evidential Perspectives*. Oxford: Oxford University Press, 247–262.

French, J. P. & Harrison, P. (2007). Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *The International Journal of Speech, Language and the Law* **14**(1), 137–144.

French, J. P., Nolan, F., Foulkes, P., Harrison, P., & McDougall, K. (2010). The UK position statement on forensic speaker comparison: A rejoinder to Rose and Morrison. *The International Journal of Speech, Language and the Law* **17**(1), 143–152.

French, J. P. & Stevens, L. (2013). Forensic speech science. In: Jones, M. J. & Knight, R.-A. (eds.). *The Bloomsbury Companion to Phonetics*. London, New Delhi, New York, Sydney: Bloomsbury, 183–197.

Frowd, C. D., Skelton, F. C., Atherton, C. J., Pitchford, M., Hepton, G., Holden, L., McIntyre, A. H. & Hancock, P. J. B. (2012). Recovering faces from memory: the distracting influence of external facial features. *Journal of Experimental Psychology: Applied* **18**(2), 224–238.

Fuchs, S., Koenig, L. L., Winkler, R. (2007). Weak clicks in German? *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, August 6–10, 2007, 449–452.

Fuchs, S., Weirich, M., Kroos, C., Fecher, N., Pape, D. & Koppetsch, S. (2010). Time for a shave? Does facial hair interfere with visual speech intelligibility? In: Fuchs, S., Hoole, P., Mooshammer, C. & Zygis, M. (eds.). *Between the Regular and the Particular in Speech and Language*. Frankfurt/M.: Peter Lang, 247–264.

Gagné, J.-P., Masterson, V., Munhall, K. G., Bilida, N. & Querengesser, C. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal of the Academy of Rehabilitative Audiology* **27**, 135–158.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* **6**(1), 110–125.

Garner, W. R. (1974). *The Processing of Information and Structure*. New York: Halsted Press.

Gerrard, G. & Thompson, R. (2011). Two million cameras in the UK. *CCTV Image Magazine*. [Online]. Winter 2010. Available from: http://www.goo.gl/I1nYhc. [Accessed: 7th May 2014].

Gick, B. & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature* **462**, 502–504.

Giles, V. (2013). *Effects of facewear on speech perception: A multimodal study*. MSc dissertation, University of York.

Gill, M. & Spriggs, A. (2005). *Assessing the Impact of CCTV*. Home Office Research Study No. 292. London: Home Office. Available from: http://www.goo.gl/gHfKCW. [Accessed: 7th May 2014].

Goecke, R. (2005). Current trends in joint audio-video signal processing: A review. *Proceedings of the 8th International Symposium on Signal Processing and Its Applications*, Sydney, Australia, August 28–31, 2005, 70–73.

Goh, W. D. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **31**(1), 40–53.

Gold, E. & French, J. P. (2011). International practices in forensic speaker comparison. *The International Journal of Speech, Language and the Law* **18**(2), 293–307.

Goldfrank, L. R. & Liverman, C. T. (eds.) (2008). *Preparing for an Influenza Pandemic: Personal Protective Equipment for Healthcare Workers*. Washington, D.C.: The National Academies Press.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **22**(5), 1166–83.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review* **105**(2), 252–279.

Gordon, M., Barthmaier, P. & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association* **32**(2), 141–174.

Gracco, V. L. & Löfqvist, A. (1994). Speech motor coordination and control: Evidence from lip, jaw, and laryngeal movements. *The Journal of Neuroscience* **14**(11), 6585–97.

Graham-Rowe, D. (2006a). Surveillance system spots violent behaviour. *New Scientist*. [Online]. 26th October 2006. Available from: http://www.goo.gl/hB5ixa. [Accessed: 7th May 2014].

Graham-Rowe, D. (2006b). Smart statistics keep eye on CCTV. *New Scientist*. [Online]. 13 November 2003. Available from: http://www.goo.gl/PvJpHE. [Accessed: 7th May 2014].

Grant, K. W. (2003). Auditory supplements to speechreading. *Paper presented at the 'Speech Dynamics by Ear, Eye, Mouth and Machine: An Interdisciplinary Workshop' (organised by The Institute for Electronics, Information and Communication Engineers (IEICE) and The Acoustical Society of Japan)*, Kyoto, Japan, June 27, 2003.

Grant, K. W. & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America* **108**(3), 1197–208.

Greenberg, H. J. & Bode, D. L. (1968). Visual discrimination of consonants. *Journal of Speech, Language, and Hearing Research* **11**(4), 869–874.

Guillemin, B. J. & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: Some preliminary findings. *The International Journal of Speech, Language and the Law* **15**(2), 193–218.

Haley, K. L., Seelinger, E., Mandulak, K. C. & Zajac, D. J. (2010). Evaluating the spectral distinction between sibilant fricatives through a talker-centered approach. *Journal of Phonetics* **38**(4), 548–554.

Halle, M., Hughes, G. W. & Radley, J.-P. A. (1957). Acoustic properties of stop consonants. *The Journal of the Acoustical Society of America* **29**(1), 107–116.

Harrington, J. (2010). Acoustic phonetics. In: Hardcastle, W. J., Laver, J. & Gibbon, F. E. (eds.). *The Handbook of Phonetic Sciences*. 2nd ed. Malden, Oxford, Chichester: Wiley-Blackwell, 81–129.

Harrington, J., Fletcher, J. & Roberts, C. (1995). Coarticulation and the accented/unaccented distinction: Evidence from jaw movement data. *Journal of Phonetics* **23**(3), 305–322.

Harvard Law Review (2004). *Constitutional Law. Free Speech. Second Circuit Upholds New York's Anti-Mask Statute against Challenge by Klan-Related Group. Church of the American Knights of the Ku Klux Klan v. Kerik, 356 F.3d*

197 (2d Cir. 2004). Harvard Law Review 117(8), 2777–84. Available from: http://www.goo.gl/HHoki9. [Accessed: 7th May 2014].

Hawkins, S. & Stevens, K. N. (1987). Perceptual and acoustical analyses of velar stop consonants. *Proceedings of the 11th International Congress of Phonetic Sciences (ICPhS)*, Tallinn, Estonia, August 1–7, 1987, 342–345.

Hay, J. & Drager, K. (2007) Sociophonetics. *Annual Review of Anthropology* **36**, 89–103.

Hayward, K. (2000). *Experimental Phonetics*. Harlow: Longman Linguistics Library.

Hazan, V., Kim, J. & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication* **52**(11–12), 996–1009.

Hazan, V. & Li, E. (2008). The effect of auditory and visual degradation on audiovisual perception of native and non-native speakers. *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)*, Brisbane, Australia, September 22–26, 1191–4.

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M. & Chunge, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America* **119**(3), 1740–51.

Heath, A. J. & Moore, K. (2011). Earwitness memory: Effects of facial concealment on the face overshadowing effect. *International Journal of Advanced Science and Technology* **33**, 131–140.

Hogan, J. (2003). Your every move will be analysed. *New Scientist*. [Online]. 12 July 2003. Available from: http://www.goo.gl/GC9euW. [Accessed: 7th May 2014].

Hollien, H. (1990). *The Acoustics of Crime: The New Science of Forensic Phonetics*. New York, London: Plenum Press.

Hollien, H. (2002). *Forensic Voice Identification*. San Diego, London: Academic Press.

Hollien, H. & Schwartz, R. (2000). Aural-perceptual speaker identification: Problems with noncontemporary samples. *The International Journal of Speech, Language and the Law* **7**(2), 199–211.

Hombert, J.-M., Ohala, J. J. & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language* **55**(1), 37–58.

House, A. S., Williams, C. E., Heker, M. H. & Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *The Journal of the Acoustical Society of America* **37**, 158–166.

Hove, I. & Dellwo, V. (2012). The effect of articulatory obstruction on temporal characteristics of speech. *Paper presented at the 2012 International Association of Forensic Phonetics and Acoustics Annual Conference (IAFPA)*, Santander, Spain, August 6–8, 2012.

Howell, T. B. *et al.* (eds.) (1810). *Cobbett's Complete Collection of State Trials and Proceedings for High Treason and other Crimes and Misdemeanors from the Earliest Period to the Present Time*, Vol. V. (Trial of William Hulet, 1185–1195). London: R. Bagshaw. Available from: http://www.goo.gl/9yc9Ip. [Accessed: 7th May 2014].

Howell, K. & Martin, A. M. (1975). An investigation of the effects of hearing protectors on vocal communication in noise. *Journal of Sound and Vibration* **41**(2), 181–196.

IJsseldijk, F. J. (1992). Speechreading performance under different conditions of video image, repetition, and speech rate. *Journal of Speech, Language, and Hearing Research* **35**(2), 466–471.

Ito, T., Gomi, H. & Honda, M. (2000). Task dependent jaw-lip coordination examined by jaw perturbation during bilabial-consonant utterances. *Proceedings of the 5th Seminar on Speech Production: Models & Data*, Kloster Seeon, Bavaria, Germany, May 1–4, 2000, 41–44.

Ito, T., Tiede, M. & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United States of America* **106**(4), 1245–8.

Jakobson, R., Fant, G. & Halle, M. (1963). *Preliminaries to Speech Analysis. The distinctive features and their correlates.* Cambridge, Ma.: The MIT Press.

Jessen, M. (2007). Speaker classification in forensic phonetics and acoustics. In: Müller, C. (ed.). *Speaker Classification I: Fundamentals, Features, and Methods*. Berlin, Heidelberg: Springer, 80–204.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass* **2**(4), 671–711.

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *The International Journal of Speech, Language and the Law* **12**(2), 174–213.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *The Journal of the Acoustical Society of America* **88**(2), 642–654.

Johnson, K. (2003). *Acoustic & Auditory Phonetics*. 2nd ed. Malden, Oxford, Melbourne, Berlin: Blackwell Publishing Ltd.

Jones, M. & Llamas, C. (2008). Fricated realisations of /θ/ in Dublin and Middlesbrough English: An acoustic analysis of plosive frication and surface fricative contrasts. *English Language and Linguistics* **12**(3), 419–443.

Jongman, A., Wayland, R. & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America* **108**(3), 1252–63.

Jordan, T. R. & Bevan, K. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance* **23**(2), 388–403.

Jordan, T. R., McCotter, M. V. & Thomas, S. M. (2000). Visual and audiovisual speech perception with color and gray-scale facial images. *Perception & Psychophysics* **62**(7), 1394–404.

Jordan, T. R. & Sergeant, P. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech* **43**(1), 107–124.

Jordan, T. R., & Thomas, S. M. (2001). Effects of horizontal viewing angle on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance* **27**(6), 1386–403.

Jordan, T. R. & Thomas, S. M. (2011). When half a face is as good as a whole: Effects of simple substantial occlusion on visual and audiovisual speech perception. *Attention, Perception, & Psychophysics* **73**(7), 2270–85.

Kamachi, M., Hill, H., Lander, K. & Vatikiotis-Bateson, E. (2003). 'Putting the face to the voice': Matching identity across modality. *Current Biology* **13**(19), 1709–14.

Kapoor, V. (2012). Extending the perception of speech intelligibility in respiratory protection. *Paper presented at the ISRP (International Society for Respiratory Protection) 16th International Conference: A global View on Respiratory Protection*, Boston, Mass., USA, September 23–27, 2012.

Kavanagh, C. M. (2013). *New consonantal acoustic parameters for forensic speaker comparison*. Ph.D. dissertation, University of York.

Keating, P. A. (1988). Underspecification in phonetics. *Phonology* **5**(2), 275–292.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *The Journal of the Acoustical Society of America* **73**(1), 322–335.

Khattab, G. & Al-Tamimi, J. (2008). Phonetic patterns of gemination in Lebanese Arabic. *Paper presented at the 2008 Colloquium of the British Association of Academic Phoneticians (BAAP)*, Sheffield, United Kingdom, March 31–April 2, 2008.

Kim, J. & Davis, C. (2003). Hearing foreign voices: Does knowing what is said affect masked visual speech detection? *Perception* **32**(1), 111–120.

Kim, J., Davis, C. & Groot, C. (2009). Speech identification in noise: Contribution of temporal, spectral, and visual speech cues. *The Journal of the Acoustical Society of America* **126**(6), 3246–57.

Kirk, D. (2013). Appearance in court. Veiled threats. *The Journal of Criminal Law* **77**(6), 459–461.

Kirkham, S. (2011). The acoustics of coronal stops in British Asian English. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)*, Hong Kong, China, August 17–21, 2011, 1102–5.

Kitano, Y., Siegenthaler, B. M. & Stoker, R. G. (1985). Facial hair as a factor in speechreading performance. *Journal of Communication Disorders* **18**(5), 373–381.

Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research* **18**(4), 686–706.

Knösche, T. R., Lattner, S., Maess, B., Schauer, M. & Friederici, A. D. (2002). Early parallel processing of auditory word and voice information. *NeuroImage* **17**, 1493–503.

Kraljic, T. & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language* **56**(1), 1–15.

Kreiman, J. & Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar voices. *Speech Communication* **10**(3), 265–275.

Kricos, P. B. (1996). Differences in visual intelligibility across talkers. In: Stork, D. G. & Hennecke, M. E. (eds.). *Speechreading by Humans and Machines: Models, Systems, and Applications*. Berlin, Heidelberg, New York: Springer, 43–55.

Kroos, C. & Dreves, A. (2008). McGurk effect persists with a partially removed visual signal. *Proceedings of the 8th International Conference on Auditory-*

Visual Speech Processing (AVSP), Tangalooma Wild Dolphin Resort, Moreton Island, Queensland, Australia, September 26–29, 2008, 55–58.

Kryter, K. D. (1946). Effects of ear protective devices on the intelligibility of speech in noise. *The Journal of the Acoustical Society of America* **18**(2), 413–523.

Künzel, H. J. (2000). Effects of voice disguise on speaking fundamental frequency. *The International Journal of Speech, Language and the Law* **7**(2), 149–179.

Künzel, H. J. (2001). Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. *The International Journal of Speech, Language and the Law* **8**(1), 80–99.

Lachs, L. & Pisoni, D. B. (2004). Cross-modal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance* **30**(2), 378–396.

Ladefoged, P. (1967). *Three Areas of Experimental Phonetics*. London: Oxford University Press.

Ladefoged, P. (1978). Expectation affects identification by listening. *Language and Speech* **21**, 373–374.

Ladefoged, P. (2003). *Phonetic Data Analysis. An Introduction to Fieldwork and Instrumental Techniques*. Malden, Oxford, Carlton: Blackwell Publishing Ltd.

Ladefoged, P. & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America* **29**(1), 98–104.

Ladefoged, P. & Disner S. F. (2012). *Vowels and Consonants*. 3rd ed. Malden, Oxford, Chichester: Blackwell Publishing Ltd.

Ladefoged, P. & Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford, Malden: Blackwell Publishers Ltd.

Lancker, D. van & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia* **25**(5), 829–834.

Lansing, C. R. & McConkie, G. W. (1994). A new method for speechreading research: Tracking observers' eye movements. *Journal of the Academy of Rehabilitative Audiology* **27**, 25–43.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Laver, J. (1995). *Principles of Phonetics*. Cambridge, New York, Melbourne: Cambridge University Press.

Laver, J. & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In: Scherer, K. R. & Giles, H. (eds.). *Social Markers in Speech*. Cambridge: Cambridge University Press, 1–31.

Lecumberri, M. L. G. & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America* **119**(4), 2445–54.

Lesner, S. A. (1988). The Talker. *The Volta Review* **90**(5), 89–98.

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* **21**, 1–36.

Lidestam, B. & Beskow, J. (2006). Visual phonemic ambiguity and speechreading. *Journal of Speech, Language, and Hearing Research* **49**(4), 835–847.

Lisker, L. & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word* **20**(3), 384–422.

Llamas, C., Harrison, P., Donnelly, D. & Watt, D. (2008). Effects of different types of face coverings on speech acoustics and intelligibility. *York Papers in Linguistics (Series 2)* **9**, 80–104.

Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about 'weapon focus'. *Law and Human Behavior* **11**(1), 55–62.

Lovitt, A. & Allen, J. B. (2006). 50 years late: Repeating Miller-Nicely 1955. *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech, ICSLP),* Pittsburgh, PA, USA, September 17–21, 2006, 2154–7.

Machač, P. & Skarnitzl, R. (2009). *Principles of Phonetic Segmentation*. Prague: Epocha Publishing House.

Macmillan, N. A. & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. 2nd ed. Mahwah, NJ, London: Lawrence Erlbaum Associates.

Maniwa, K., Jongman, A. & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America* **125**(6), 3962–73.

Mann, S. (2013). Wearable computing. In: Soegaard, M. & Dam, R. F. (eds.). *The Encyclopedia of Human-Computer Interaction*. 2nd ed. Aarhus, Denmark: The

Interaction Design Foundation. Available from: http://www.goo.gl/iY65th. [Accessed: 7th May 2014].

Mann, S., Nolan, J. & Wellman, B. (2003). Sousveillance: Inventing and using wearable computing devices for data collection in surveillance environments. *Surveillance & Society* **1**(3), 331–355.

Marassa, L. K. & Lansing, C. R. (1995). Visual word recognition in two facial motion conditions: Full-face versus lips-plus-mandible. *Journal of Speech, Language, and Hearing Research* **38**(6), 1387–94.

Martin, A. M., Howell, K. & Lower, M. C. (1976). Hearing protection and communication in noise. *Proceedings of the 2nd Conference of the British Society of Audiology*, Southampton, United Kingdom, July 16–18, 1975, 47–62.

Mason, J. S. D., Deravi, F., Chibelushi, C. C. & Gandon, S. (1996). *Project: DAVID (Digital Audio Visual Integrated Database)* (Final Report, Contract Number ML649592, 26 September 1996). Swansea, United Kingdom: Department of Electrical and Electronic Engineering, University of Wales Swansea. Available from: http://www.goo.gl/XPHMjK. [Accessed: 7th May 2014].

Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, Mass.: MIT Press.

Massaro, D. W. (2001). Speech perception. In: Smelser, N. M., Baltes, P. B. & Kintsch, W. (eds.). *International Encyclopedia of Social and Behavioral Sciences*. 2nd ed. Amsterdam: Elsevier, 14870–5.

McAllister, H. A., Dale, R. H. I., Bregman, N. J., McCabe, A. & Cotton, C. R. (1993). When eyewitnesses are also earwitnesses: Effects on visual and voice identifications. *Basic and Applied Social Psychology* **14**(2), 161–170.

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology* **18**, 1–86.

McDougall, K. & Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, August 6–10, 2007, 1825–8.

McFarland, D. H. & Baum, S. R. (1995). Incomplete compensation to articulatory perturbation. *The Journal of the Acoustical Society of America* **97**(3), 1865–73.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* **264**, 746–748.

Ménard, L., Perrier, P. & Aubin, J. (2013). The role of auditory feedback in speech development: A study of compensation strategies for a lip-tube perturbation. *The Journal of the Acoustical Society of America* **133**(5), 3564–3564.

Mendel, L. L., Gardino, J. A. & Atcherson, S. R. (2008). Speech understanding using surgical masks: A problem in health care? *Journal of the American Academy of Audiology* **19**(9), 686–695.

Mendoza-Denton, N. C. (1997). *Chicana/Mexicana identity and linguistic variation: An ethnographic and sociolinguistic study of gang affiliation in an urban high school*. Ph.D. dissertation, Stanford University.

Milenkovic, P. (1986). Glottal inverse filtering by joint estimation of an AR system with a linear input model. *IEEE Transactions on Acoustics, Speech and Signal Processing* **34**(1), 28–42.

Miller, G. A. & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America* **27**(2), 338–352.

Miller, J. L. (1978). Interactions in processing segmental and suprasegmental features of speech. *Perception & Psychophysics* **24**(2), 175–180.

Morrow, C. T. (1947). Reaction of small enclosures on the human voice, part II: Analyses of vowels. *The Journal of the Acoustical Society of America* **19**(4), 487–497.

Mullennix, J. W. & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics* **47**(4), 379–390.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T. & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science* **15**(2), 133–137.

Munhall, K. G., Kroos, C., Jozan, G. & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics* **66**(4), 574–583.

Munhall, K. G. & Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In: Campbell, R., Dodd, B. & Burnham, D. (eds.). *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*. Hove: Psychology Press Ltd, 123–139.

Munhall, K. G. & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In: Calvert, G. A., Spence, C. & Stein, B. E. (eds.). *The Handbook of Multisensory Processes*. Cambridge, Mass.: The MIT Press, 177–188.

Munson, B. (2001). A method for studying variability in fricatives using time-dependent changes in spectral mean. *The Journal of the Acoustical Society of America* **109**(5), 2293–2293.

Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge, London, New York: Cambridge University Press.

Nolan, F. (1997). Speaker recognition and forensic phonetics. In: Hardcastle, W. J. & Laver, J. (eds.). *The Handbook of Phonetic Sciences*. Oxford: Blackwell, 744–767.

Nolan, F. (2001). Speaker identification evidence: Its forms, limitations and roles. *Paper presented at the 'Law and Language: Prospect and Retrospect' conference*, Levi, Finnish Lapland, December 12–15, 2001.

Nolan, F. (2003). A recent voice parade. *The International Journal of Speech, Language and the Law* **10**(2), 277–291.

Nolan, F. (2012). Degrees of freedom in speech production: An argument for native speakers in LADO. *The International Journal of Speech, Language and the Law* **19**(2), 263–289.

Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *The International Journal of Speech, Language and the Law* **12**(2), 143–173.

Noy, D. (2003). *Evaluation of transmission loss induced by stretched fabric treatments*. [Online]. Available from: http://www.goo.gl/QwCYfm. [Accessed: 7th May 2014].

Nute, M. E. & Slater, K. (1973). The effect of fabric parameters on sound transmission loss. *The Journal of the Textile Institute* **64**(11), 652–658.

Nygaard, L. C. & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics* **60**(3), 355–376.

Nygaard, L. C., Sommers, M. S. & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science* **5**(1), 42–46.

Ofcom (2013). *Communications Market Report 2013* (1 August 2013). London: Ofcom. Available from: http://www.goo.gl/0WBsD4. [Accessed: 7th May 2014].

Onaka, A. & Watson, C. I. (2000). Acoustic comparison of child and adult fricatives. *Proceedings of the 8th Australian International Conference on Speech Science and Technology*, Canberra, Australia, December 5–7, 2000, 134–139.

Oxford English Dictionary (2013). 2nd ed. Oxford: Oxford University Press. [Online]. Available from: http://www.goo.gl/EmKOTq [Accessed: 7th May 2014].

Pääkkönen, R., Lehtomäki, K., Savolainen, S., Myllyniemi, J. & Hämäläinen, E. (2000). Noise attenuation of hearing protectors against heavy weapon noise. *Military Medicine* **165**(9), 678–682.

Paré, M., Richler, R. C., ten Hove, M. & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics* **65**(4), 553–567.

Patrick, P. L. (2012). Language analysis for the determination of origin: Objective evidence for refugee status determination. In: Tiersma, P. M. & Solan, L. M. (eds.). *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press, 533–546.

Pepiot, E. (2012). Voice, speech and gender: male-female acoustic differences and cross-language variation in English and French speakers. *Actes des XVèmes Rencontres Jeunes Chercheurs de l'ED 268, Paris, France, 2011-2012*.

Peláez-Moreno, C., García-Moral, A. I., Valverde-Albacete, F. J. (2010). Analyzing phonetic confusions using formal concept analysis. *The Journal of the Acoustical Society of America* **128**(3), 1377–90.

Perfect, T. J., Hunt, L. J. & Harris, C. M. (2002). Verbal overshadowing in voice recognition. *Applied Cognitive Psychology* **16**(8), 973–980.

Phatak, S. A. & Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America* **121**(4), 2312–26.

Phatak, S. A., Lovitt, A. & Allen, J. B. (2008). Consonant confusions in white noise. *The Journal of the Acoustical Society of America* **124**(2), 1220–33.

Pickel, K. L., French, T. A. & Betts, J. M. (2003). A cross-modal weapon focus effect: The influence of a weapon's presence on memory for auditory information. *Memory* **11**(3), 277–292.

Pisoni, D. B. (1997). Some thoughts on 'normalization' in speech perception. In: Johnson, K. & Mullennix, J. W. (eds.). *Talker Variability in Speech Processing*. San Diego: Academic Press, 9–32.

Preminger, J. E., Lin, H.-B., Payen, M. & Levitt, H. (1998). Selective visual masking in speechreading. *Journal of Speech, Language and Hearing Research* **41**(3), 564–575.

Radonovich, L. J. Jr., Yanke, R., Cheng, J. & Bender, B. (2010). Diminished speech intelligibility associated with certain types of respirators worn by healthcare workers. *Journal of Occupational and Environmental Hygiene* **7**(1), 63–70.

Redford, M. A. & Diehl, R. L. (1999). The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *The Journal of the Acoustical Society of America* **106**(3), 1555–65.

Remez, R. E., Fellowes, J. M. & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance* **23**(3), 651–666.

Rife, D. & Vanderkooy, J. (1989). Transfer function measurement with Maximum-Length Sequences. *Journal of the Audio Engineering Society* **37**(6), 419–443.

Roach, P. (2004). British English: Received Pronunciation. *Journal of the International Phonetic Association* **34**(2), 239–245.

Roberge, R. J. (2008). Effect of surgical masks worn concurrently over N95 filtering facepiece respirators: Extended service life versus increased user burden. *Journal of Public Health Management and Practice* **14**(2), E19–26.

Rönnberg, J., Samuelsson, S. & Lyxell, B. (1998). Conceptual constraints in sentence-based lipreading in the hearing impaired. In: Campbell, R., Dodd, B. & Burnham, D. (eds.). *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*. Hove: Psychology Press Ltd, 143–153.

Rose, P. (2002). *Forensic Speaker Identification*. London, New York: Taylor & Francis.

Rose, P. & Morrison, G. S. (2009). A response to the UK Position Statement on forensic speaker comparison. *The International Journal of Speech, Language and the Law* **16**(1), 139–163.

Rosenblum, L. D. (2005). Primacy of multimodal speech perception. In: Pisoni, D. B. & Remez, R. E. (eds.). *The Handbook of Speech Perception*. Blackwell Publishing Ltd, 51–78.

Rosenblum, L. D., Johnson, J. A. & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research* **39**(6), 1159–70.

Rosenblum, L. D., Miller, R. M. & Sanchez, K. (2007). Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. *Psychological Science* **18**(5), 392–396.

Rosenblum, L. D. & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance* **22**(2), 318–331.

Rosenblum, L. D., Yakel, D. A. & Green, K. P. (2000). Face and mouth inversion effects on visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance* **26**(2), 806–819.

Salazar, M. K., Connon, C., Takaro, T. K., Beaudet, N. & Barnhart, S. (2001). An evaluation of factors affecting hazardous waste workers' use of respiratory protective equipment. *American Industrial Hygiene Association Journal* **62**(2), 236–245.

Scarborough R., Keating, P., Mattys, S. L., Cho, T. & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech* **51**(2/3), 135–175.

Scheinberg, J. S. (1980). Analysis of speechreading cues using an interleaved technique. *Journal of Communication Disorders* **13**(6), 489–492.

Schiel, F. (2004). MAUS goes iterative. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, May 26–28, 2004, 1015–8.

Schiller, N. O. & Köster, O. (1998). The ability of expert witnesses to identify voices: A comparison between trained and untrained listeners. *The International Journal of Speech, Language and the Law* **5**(1), 1–9.

Schiller, N. O., Köster, O. & Duckworth, M. (1997). The effect of removing linguistic information upon identifying speakers of a foreign language. *The International Journal of Speech, Language and the Law* **4**(1), 1–17.

Schooler, J. W. & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology* **22**(1), 36–71.

Schwartz, J.-L., Berthommier, F. & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition* **93**(2), B69–78.

Sekiyama, K. & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science* **11**(2), 306–320.

Shadle, C. (1985). Intrinsic fundamental frequency of vowels in sentence context. *The Journal of the Acoustical Society of America* **78**(5), 1562–7.

Shadle, C. & Mair, S. J. (1996). Quantifying spectral characteristics of fricatives. *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, PA, USA, October 3–6, 1996, 1521–4.

Shahin, A. J. & Miller, L. M. (2009). Multisensory integration enhances phonemic restoration. *The Journal of the Acoustical Society of America* **125**(3), 1744–50.

Sheffert, S. M. & Olson, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics* **66**(2), 352–362.

Simoni, S. J. (1992). 'Who goes there?' – Proposing a model anti-mask act. *Fordham Law Review* **61**(1), Article 16, 241–247.

Simpson, S. A. & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *The Journal of the Acoustical Society of America* **118**(3), 2775–8.

Smits, R. (2000). Temporal distribution of information for human consonant recognition in VCV utterances. *Journal of Phonetics* **28**(2), 111–135.

Smits, R., Warner, N., McQueen, J. M. & Cutler, A. (2003). Unfolding of phonetic information over time: A database of Dutch diphone perception. *The Journal of the Acoustical Society of America* **113**(1), 563–574.

Soli, S. D. & Arabie, P. (1979). Auditory versus phonetic accounts of observed confusions between consonant phonemes. *The Journal of the Acoustical Society of America* **66**(1), 46–59.

Solan, L. M. & Tiersma, P. M. (2005). *Speaking of Crime. The Language of Criminal Justice*. Chicago, London: The University of Chicago Press.

Sommer, H. C. (1976). *Speech communication capability and hearing protection of USAF inflight headgear devices* (Technical Report AMRL-TR-75-67, June 1976). Springfield, VA, USA: Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio 45433. Available from: http://www.goo.gl/65vdpK. [Accessed: 7th May 2014].

Stanton, B. J., Jamieson, L. H. & Allen, G. D. (1988). Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New York, NY, USA, April 11–14, 1988, 331–334.

Stephens, J. D. W. & Holt, L. L. (2010). Learning to use an artificial visual cue in speech identification. *The Journal of the Acoustical Society of America* **128**(4), 2138–49.

Stevenage, S. V., Howland, A. & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology* **25**(1), 112–118.

Stevenage, S. V., Neil, G. J., Barlow, J., Dyson, A., Eaton-Brown, C. & Parsons, B. (2012). The effect of distraction on face and voice recognition. *Psychological Research* **77**(2), 167–175.

Stevenage, S. V., Neil, G. J. & Hamlin, I. (2013). When the face fits: Recognition of celebrities from matching and mismatching faces and voices. *Memory* **22**(3), 284–294.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In: David, E. E. & Denes, P. B. (eds.). *Human Communication: A Unified View*. New York: McGraw-Hill, 51–66.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics* **17**, 3–46.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, Mass., London: The MIT Press.

Stevens, K. N. (2010). Articulatory-acoustic relations as the basis of distinctive contrasts. In: Hardcastle, W. J., Laver, J. & Gibbon, F. E. (eds.). *The Handbook of Phonetic Sciences*. 2nd ed. Malden, Oxford, Chichester: Wiley-Blackwell, 425–453.

Stevens, M. & Hajek, J. (2004). A preliminary investigation of some acoustic characteristics of ejectives in Waima'a: VOT and closure duration. *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, Sydney, Australia, December 8–10, 2004, 277–282.

Stevens, K. N. & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics* **38**(1), 10–19.

Stoel-Gammon, C., Williams, K. & Buder, E. (1994). Cross-language differences in phonological acquisition: Swedish and American /t/. *Phonetica* **51**(1–3), 146–158.

Stone, L. (1957). *Facial Cues of Context in Lip Reading*. John Tracy Clinic Research Papers 5, Los Angeles: John Tracy Clinic.

Stuart-Smith, J., Timmins, C. & Wrench, A. (2003). Sex and gender in Glaswegian /s/. *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, August 3–9, 2003, 1851–4.

Süddeutsche Zeitung (2010). Der Kopftuch-Streit in Deutschland. [Online]. 19th May 2010. Available from: http://www.goo.gl/63qq18. [Accessed: 7th May 2014].

Sumby, W. H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America* **26**(2), 212–215.

Summerfield, A. Q. (1979). Use of visual information for phonetic perception. *Phonetica* **36**(4–5), 314–331.

Summerfield, A. Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd, B. & Campbell, R. (eds.). *Hearing by Eye: The Psychology of Lip-Reading*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

Sundara, M. (2005). Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French. *The Journal of the Acoustical Society of America* **118**(2), 1026–37.

Swerts, M. & Krahmer, E. (2006). The importance of different facial areas for signalling visual prominence. *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech, ICSLP)*, Pittsburgh, PA, USA, September 17–21, 2006, 280–283.

Swerts, M. & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics* **36**(2), 219–238.

Tabain, M. (1998). Non-sibilant fricatives in English: Spectral information above 10kHz. *Phonetica* **55**(3), 107–130.

Tabain, M. & Watson, C. (1996). Classification of fricatives. *Proceedings of the 6th Australian International Conference on Speech Science and Technology*, Adelaide, Australia, December 10–12, 1996, 623–628.

The World Post (2013). Veil bans by country: A look at restrictions on Muslim headscarves around the world (photos). [Online]. 18th September 2013. Available from: http://www.goo.gl/wmEFmi. [Accessed: 7th May 2014].

Thomas, S. M. & Jordan, T. R. (2002). Determining the influence of Gaussian blurring on inversion effects with talking faces. *Perception & Psychophysics* **64**(6), 932–944.

Thomas, S. M. & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance* **30**(5), 873–888.

Tiippana, K., Andersen, T. S. & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology* **16**(3), 457–472.

Todd, S. (2010). The 'veiling' question: On the demand for visibility in communicative encounters in education. *Philosophy of Education* 2010, 349–356.

Tomiak, G. R. (1990). An evaluation of a spectral moments metric with voiceless fricative obstruents. *The Journal of the Acoustical Society of America* **87**(S1), S106–S106.

Traunmüller, H. & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America* **107**(6), 3438–51.

Trojanová, J., Hrúz, M., Campr, P. & Železný, M. (2008). Design and recording of Czech audio-visual database with impaired conditions for continuous speech recognition. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 28–30, 2008, 1239–43.

Tubbs, R. L. (1995). Noise and hearing loss in firefighting. *Occupational Medicine* **10**(4), 843–856.

Tufts, J. B. & Frank, T. (2003). Speech production in noise with and without hearing protection. *The Journal of the Acoustical Society of America* **114**(2), 1069–80.

Tuomainen, J., Andersen, T. S., Tiippana, K. & Sams, M. (2005). Audio-visual speech perception is special. *Cognition* **96**(1), B13–22.

Turk, A., Nakai, S. & Sugahara, M. (2006). Acoustic segment durations in prosodic research: A practical guide. In: Sudhoff, S. *et al.* (eds.). *Methods in Empirical Prosody Research*. Berlin, New York: Walter de Gruyter, 1–28.

UK Judicial Studies Board (2013). *Equal Treatment Bench Book*. Available from: http://www.goo.gl/635rZJ. [Accessed: 7th May 2014].

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S. & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics* **60**(6), 926–940.

Vatikiotis-Bateson, E. & Ostry, D. J. (1999). Analysis and modeling of 3D jaw motion in speech and mastication. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Tokyo, Japan, October 12–15, 1999, 442–447.

Vicenik, C. (2010). An acoustic study of Georgian stop consonants. *Journal of the International Phonetic Alphabet* **40**(1), 59–92.

Vojnovic, M. & Mijić, M. (1997). The influence of the oxygen mask on long-time spectra of continuous speech. *The Journal of the Acoustical Society of America* **102**(4), 2456–8.

Wagoner, L., McGlothlin, J., Chung, K., Strickland, E., Zimmerman, N. & Carlson, G. (2007). Evaluation of noise attenuation and verbal communication capabilities using three ear insert hearing protection systems among airport maintenance personnel. *Journal of Occupational and Environmental Hygiene* **4**(2), 114–122.

Wang, M. D. & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America* **54**(5), 1248–66.

Warner, N. (1996). Acoustic characteristics of ejectives in Ingush. *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, PA, USA, October 3–6, 1996, 1525–8.

Watt, D. (2010). The identification of the individual through speech. In: Llamas, C. & Watt, D. (eds.). *Language and Identities*. Edinburgh: Edinburgh University Press, 76–85.

Weber, A. & Smits, R. (2003). Consonant and vowel confusion patterns by American English listeners. *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, August 3–9, 2003, 1437–40.

Welsh, B. C. & Farrington, D. P. (2007). *Closed-circuit television surveillance and crime prevention. A systematic review* (October 2007). Stockholm: Swedish Council for Crime Prevention, Information and publications. Available from: http://www.goo.gl/w86J1k. [Accessed: 7th May 2014].

Whalen, D. H. & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics* **23**, 349–366.

Wijngaarden, S. J. van & Rots, G. (2001). Balancing speech intelligibility versus sound exposure in selection of personal hearing protection equipment for Chinook aircrews. *Aviation, Space, and Environmental Medicine* **72**(11), 1037–44.

Wilde, G. & Humes, L. E. (1990). Application of the articulation index to the speech recognition of normal and impaired listeners wearing hearing protection. *The Journal of the Acoustical Society of America* **87**(3), 1192–9.

Wilding, J., Cook, S. & Davis, J. (2000). Sound familiar? *The Psychologist* **13**(11), 558–562.

Winet, E. D. (2012). Face-veil bans and anti-mask laws: State interests and the right to cover the face. *Hastings International and Comparative Law Review* **35**(1), 217–252.

Winters, S. J., Levi, S. V. & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America* **123**(6), 4524–38.

Witkin, B. R. (1990). Listening theory and research: The state of the art. *International Journal of Listening* **4**(1), 7–32.

Wittum, K. J., Feth, L. L. & Hoglund, E. M. (2013). The effects of surgical masks on speech perception in noise. *The Journal of the Acoustical Society of America* **133**(5), 3391–3391.

Woods, D. L., Yund, E. W., Herron, T. J. & Ua Cruadhlaoich, M. A. (2010). Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise. *The Journal of the Acoustical Society of America* **127**(3), 1609–23.

Yakel, D. A., Rosenblum, L. D. & Fortier, M. A. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics* **62**(7), 1405–12.

Yarmey, A. D. (2003). Earwitness identification over the telephone and in field settings. *The International Journal of Speech, Language and the Law* **10**(1), 62–74.

Yarmey, A. D. (2004). Common-sense beliefs, recognition and the identification of familiar and unfamiliar speakers from verbal and non-linguistic vocalizations. *The International Journal of Speech, Language and the Law* **11**(2), 267–277.

Yarmey, A. D. (2012). Factors affecting lay persons' identification of speakers. In: Tiersma, P. M. & Solan, L. M. (eds.). *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press, 547–556.

Yehia, H., Rubin, P. & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication* **26**, 23–43.

Zhang, C. & Tan, T. (2008). Voice disguise and automatic speaker recognition. *Forensic Science International* 175(2–3), 118–122.