Visual Speech Enhancement and its Application in Speech Perception Training

Najwa M. Alghamdi
The Department of Computer Science
University of Sheffield

A thesis submitted for the degree of Doctor of Philosophy

September 2017

Supervisors: Dr Steve Maddock, Dr Jon Barker and Prof Guy J. Brown

To Abdulwahab, Azzah and Musfer – the sources of my strength
To Deema and Yara – my reasons to keep going

Declaration

I hereby declare that I am the sole author of this thesis. The contents of this thesis are my original work and have not been submitted for any other degree or any other university. Parts of the work presented in Chapters 4, 5 and 6 have been published in a journal and conference proceedings as follows:

1. Alghamdi, N., Maddock, S., Brown, G.J. and Barker, J. 'Investigating the impact of artificial enhancement of lip visibility on the intelligibility of spectrally-distorted speech'. 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP), 11–13 September, 2015, Vienna, Austria.
2. Alghamdi, N., Maddock, S., Brown, G.J. and Barker, J. 'A comparison of audiovisual and auditory-only training on the perception of spectrally-distorted speech'. International Congress of Phonetic Sciences (ICPhS), 10–14 August 2015, Glasgow, UK.
3. Alghamdi, N., Maddock, S., Barker, J., Brown, G. J. (2017). The impact of automatic exaggeration of the visual articulatory features of a talker on the intelligibility of spectrally distorted speech. *Speech Communication*, 95, 127-136.

Abstract

This thesis investigates methods for visual speech enhancement to support auditory and audiovisual speech perception. Normal-hearing non-native listeners receiving cochlear implant (CI) simulated speech are used as 'proxy' listeners for CI users, a proposed user group who could benefit from such enhancement methods in speech perception training. Both CI users and non-native listeners share similarities with regards to audiovisual speech perception, including increased sensitivity to visual speech cues.

Two enhancement methods are proposed: (i) an appearance based method, which modifies the appearance of a talker's lips using colour and luminance blending to apply a 'lipstick effect' to increase the saliency of mouth shapes; and (ii) a kinematics based method, which amplifies the kinematics of the talker's mouth to create the effect of more pronounced speech (an 'exaggeration effect'). The application that is used to test the enhancements is speech perception training, or audiovisual training, which can be used to improve listening skills.

An audiovisual training framework is presented which structures the evaluation of the effectiveness of these methods. It is used in two studies. The first study, which evaluates the effectiveness of the lipstick effect, found a significant improvement in audiovisual and auditory perception. The second study, which evaluates the effectiveness of the exaggeration effect, found improvement in the audiovisual perception of a number of phoneme classes; no evidence was found of improvements in the subsequent auditory perception, as audiovisual recalibration to visually exaggerated speech may have impeded learning when used in the audiovisual training.

The thesis also investigates an example of kinematics based enhancement which is observed in Lombard speech, by studying the behaviour of visual Lombard phonemes in different contexts. Due to the lack of suitable datasets for this analysis, the thesis presents a novel audiovisual

Lombard speech dataset recorded under high SNR, which offers two, fixed head-pose, synchronised views of each talker in the dataset.

Acknowledgements

During the last four years, I have met numerous challenges and accomplished my goals through the invaluable support and guidance provided by many people in both my personal and professional lives.

This thesis would not have been possible without the immense support and patience offered by my first supervisor Dr Steve Maddock, who both encouraged and supported my ideas, which made me a better person, student and researcher. His insight and advice have sustained me throughout every stage of my PhD journey, including the struggle to manage the competing demands of my studies and my family life. He went above and beyond to read every line of my drafts in detail. I have learnt more from him than I could ever express. Dr Maddock is the true definition of a mentor.

I would also like to thank my secondary supervisors Dr Jon Barker and Prof Guy J. Brown for their constructive comments and suggestions regarding my work and publications. I would like to especially thank Dr Barker for his help and support during the dataset recording.

I was honoured to work with this exceptional supervisory team, from whom I learnt many things and to whom I looked as role models. My hope is that we will have the opportunity to collaborate on future work.

I also like to thank my examiners, Dr Stuart Cunningham and Dr Ben Milner for the valuable comments and an intellectually stimulating and enjoyable viva.

My sincere thanks go to my colleagues in the Visual Computing Lab and the Speech and Hearing Lab for their continuous assistance, especially Robert Chisholm, Peter Heywood, Matthew Leach, John Charlton, Saeed Makram and Erfan Lowaimi for their invaluable help in my recording experiment.

My friends also deserve many thanks for their incredible support, particularly Mozhgan, Rabab, Sadeen, Nazrina, Aysha, Eidah and Mashael.

I give many thanks to my family for keeping me strong throughout this journey. Thank you to my grandfather Saeed for empowering the women in my family, to my aunt Samha and my cousins for arranging camping trips to the Saudi deserts whenever I am home, and to my nephews (my mosqueteros) Waleed and Mohmmad and my adorable nieces Layla, Nora and Farah for bringing joy to the family. I also feel incredibly blessed for the love and support that I continuously receive from my brother and sisters. Thank you to my beloved brother Ahmad, who never hesitates to offer help, even with his busy schedule, my kind sister-in-law Kholoud for her warm welcomes, my loving sister Sara for her loving actions, my caring sister Waad for her 'around the clock' medical help and advice and my baby sister Bushra for her incredible sense of humour.

My dear parents Azzah and Musfer, thanking you is an understatement. I would be neither who nor what I am without you.

أمي هي مأمني وأماني, وأبي هو استقامة ظهري

My princess Deema, thank you for taking care of your sister when I was not around. Because of you, I was able to balance my work with our home life. My princess Yara, your jokes and giggles after a long day of work made a big difference. My beloved husband Abdulwahab, your unconditional and genuine love, support and understanding underpinned my persistence in this journey and made the completion of this thesis possible.

Finally, the work reported in this thesis would not have been possible without the financial support of King Saud University and the Saudi Ministry of Education, to which I am grateful. The recording experiment was also funded by the UK Engineering and Physical Sciences Research Council (EPSRC project AV-COGHEAR, EP/M026981/1).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The robustness of human speech perception arises from a listener's ability to integrate and evaluate information from multiple sources. 'Audiovisual integration' refers to a listener's ability to utilise auditory and visual speech information in order to interpret the perceived message from the talker [204, 267]. The illusion of perceiving a new audio signal when listeners are presented with an incongruent audiovisual signal – known as the McGurk effect [211] – provides compelling evidence of the synergy of audio and visual speech during perception. The talker's face is a mine of valuable information: the external articulators (i.e., the lips, teeth and tongue) can provide a significant proportion of the overall visual speech information gathered from the face [210, 297]. This complementary support is also evident in adverse listening conditions, such as when listening to speech in noisy conditions, where visual speech cues can improve speech intelligibility by 5–22 dB [85, 198, 221, 296].

Speech intelligibility, or the extent of how understandable the speech is, can deteriorate for several reasons [19, 208]: external factors, such as competing sound sources and reverberation [55], or internal factors, where listeners suffer limitations in perceptual skills, such as in the case of non-native listeners and cochlear implant users. When listening to native speech, non-native listeners' perception deteriorates because of several internal factors, including the non-native speaker's limited experience with the native language [93, 294, 319]. Although cochlear implants have revolutionised the treatment of sensorineural hearing impairment, the amount of acoustic information CI users can receive is a function of various physical and physiological factors, including the number of implanted and activated electrodes and the severity of damage to the inner ear and the auditory nerve [237, 246]. Such variability in non-native listeners' and CI users' perceptions may be regarded as a source of internal adversity.

The perception of both non-native listeners [106, 175, 209, 319] and CI users [185, 226, 246] suffers increased deterioration under external noise conditions.

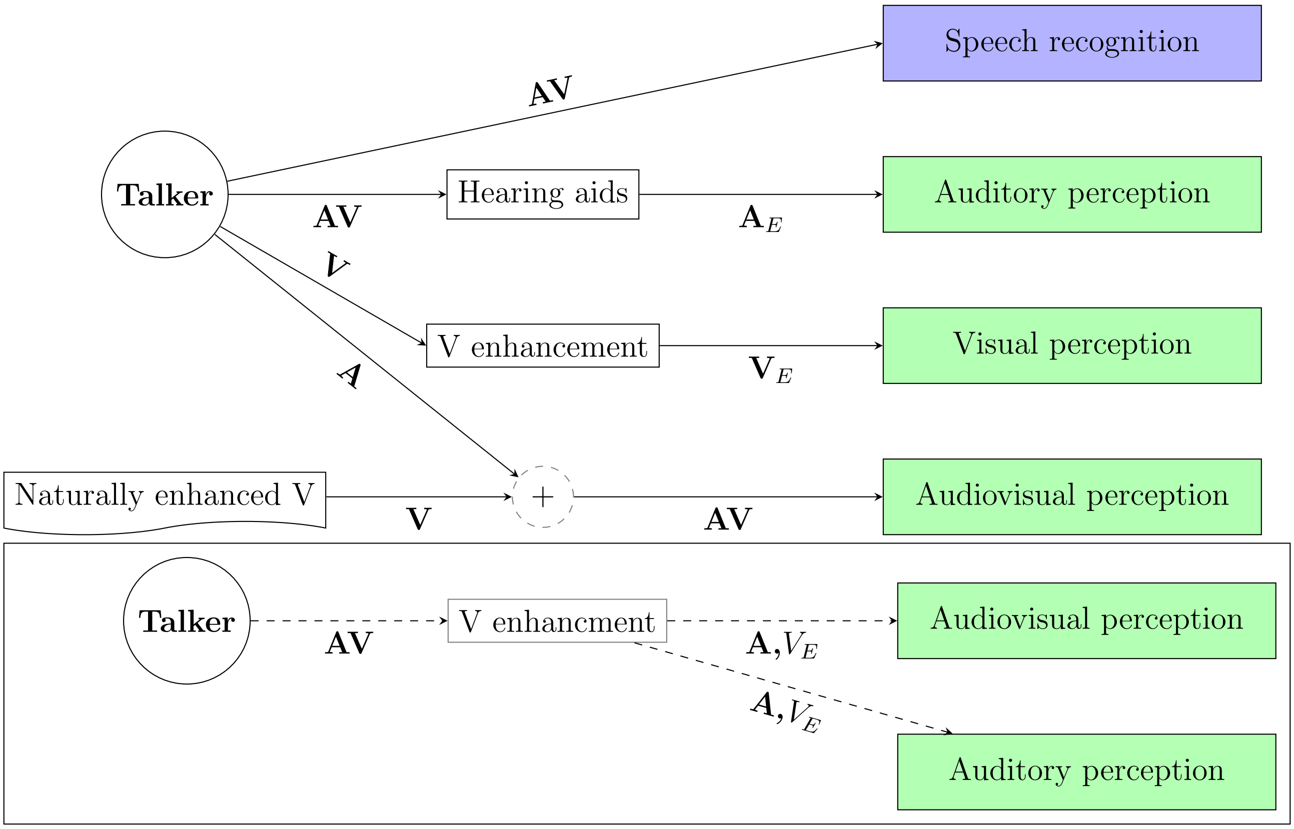


Figure 1.1: Schematic view of using visual speech in speech enhancement. Category 1 (purple): visual speech is used to enhance recognition by machines. Category 2 (green): visual speech is used to enhance perception by humans. The bottom rectangle: the gap in the literature that the thesis addresses. Subscript E denotes 'enhanced', and *A, V, AV* denote audio, visual and audiovisual speech, respectively. Naturally enhanced V: is hyper-articulated speech produced by the talker.

Fortunately, those who suffer internal adversity are known to be better audiovisual speech integrators: non-native listeners [130–132] and CI users [153, 261, 295, 317] alike are both very sensitive to visual speech cues, for example, and their perception performance generally improves when they can see the talker's face [74, 80, 130].

The significant role of visual speech in speech perception has inspired a body of work that has utilised visual speech features for speech enhancement. Figure 1.1 classifies the main categories in the speech-enhancement literature that address the use of visual speech. The first category (in purple) is the employment of visual speech features to enhance speech recognition by machine [8, 16, 251, 252]; the second category (in green), which has generally attracted less interest to date, uses visual speech to enhance human perception, which is the focus of this thesis.

The second category includes the development of hearing-assistive technology that can 'see' [7, 143] – in other words, can improve the function of hearing aids by incorporating visual cues that enhance the intelligibility of the speech signal. It also includes enhancing the visual speech signal in videos to support lip-reading [309] and enhancing audiovisual perception by combining the auditory speech signal with a simulation of previously collected visual speech data that is naturally enhanced by the talker [13] – in this case, the talker produces hyper-articulated speech to increase speech intelligibility.

1.1 Thesis Aims

The role of visual speech in supporting the auditory and audiovisual perception of those undergoing internal sources of adversity has provided the motivation to investigate and propose several methods for visual speech enhancement in this thesis. The proposed listener group of this enhancement are CI users. The best environments in which CI users prefer to interact are those that provide both auditory and visual speech signals [141]. A recent survey (Figure 1.2) conducted by Dorman *et al.* [77] on 131 CI users (61% using bilateral CIs; 31.7% using single CI) has confirmed this finding: the study showed that CI users preferred communication settings in which the talker's face was available (i.e., visual speech). In this thesis, the visual speech is enhanced to support the target listener's auditory and audiovisual perception. The enhanced visual speech is combined with the original auditory speech signal to create visually enhanced audiovisual speech.

One obstacle that CI researchers face is gaining access to a homogeneous user group. The variability in CI outcomes observed in users complicates the finding of a controlled group for testing [208, 281]. One approach some researchers have used is to use a simulation of how a CI processes speech and then to present the simulation to a normal-hearing listener as a listening model that predicts CI users' listening [78, 282]. Such models, however, do not consider the internal masking that CI users cope with [65, 208]. For this reason, normal-hearing non-native listeners have been used in this thesis as listeners who might better predict the performance of CI users as both listener groups share the effect of internal adversity [19, 106, 175, 208, 209, 225, 226, 237, 246, 319], and a sensitivity to visual cues [74, 77, 132, 153, 261, 293].

The proposed enhancement methods in this thesis are based on natural effects that have been found to be effective in supporting a listener's visual perception. For example, talkers become more visually intelligible when they wear lipstick, since it adds more definition of the mouth's shapes during speech production [173].

Talkers also tend to change their speaking style by hyper-articulating (i.e., increasing their articulation and vocal effort) to aid in communication [187]. Talkers may enunciate more when their listeners undergo a source of adversity, such as among hearing-impaired listeners or non-native listeners.

Inspired by these effects, two methods of visual speech enhancement are proposed in this study. The first method, 'appearance based enhancement', creates a realistic lipstick effect on a talker's lips in a video using colour and luminance-blending techniques. In the second method, 'kinematics based enhancement', the kinematics, or motion, of the talker's mouth movement is exaggerated in a video to create an enunciation effect on the produced speech. This is achieved by amplifying the talker's mouth shapes (using an approach based on the Principal Components Analysis) and then re-animating the video using image-warping techniques. These proposed enhancement methods modify the talker's visual speech data without using any examples of intrinsically enhanced visual speech data that the talker has made to guide the automatic enhancement, as in [13]. Hence, these enhancement methods can be applied to any communication setting that involves the presentation of a talker's face.

The application that is chosen to test the visual speech enhancement method is audiovisual or auditory training, in which the talker's face is presented; this is a speech perception training that CI users undertake to improve their listening skills. Recent evidence has shown that audiovisual training can create long-lasting improvements in subsequent auditory listening skills after the training [28, 250]. The visual speech signal guides effective perceptual learning [98, 233, 312] which in turn induces the re-organisation of the central auditory system's neural map and then enhances its response to auditory stimuli [28]. This thesis's hypothesis is that using visually enhanced audiovisual speech as training stimuli in audiovisual training may increase visual speech support during the training (thus enhancing audiovisual perception), thereby improving the post-training auditory-only skills (thus enhancing auditory training). Figure 1.1 illustrates the gap in the literature this thesis addresses.

This thesis also explores an example of natural enhancement in visual speech by investigating the visual modifications observed in speech produced in the 'Lombard' effect [192], which is an unconscious reaction that is regulated by self-monitoring of the voice. The effect is typically induced by a noise masker that results in a talker being unable to hear his or her own voice. As a response, talkers reflexively increase their vocal effort. The Lombard effect is also driven by the need to maintain intelligible communication during noisy conditions. Talkers in this case respond by

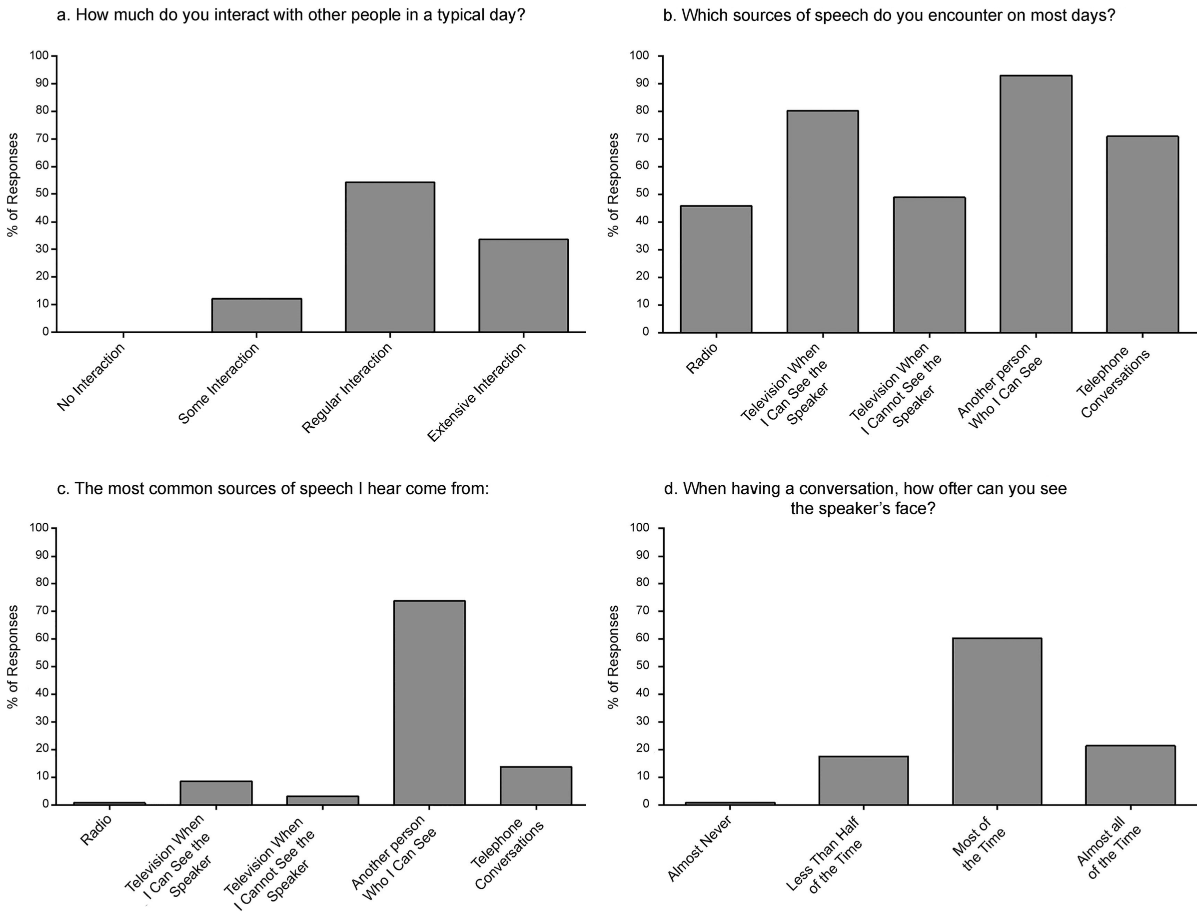


Figure 1.2: Results of a survey about listening environments that CI users experience (The figure is courtesy of Dorman *et al.* [77].)

increasing their intelligibility to aid in communication [193]. Lombard speech involves a set of acoustic and phonetic modifications, including an increase in fundamental frequency (F0), speech level and word duration [152], stronger articulatory gestures, including mouth aperture and rounding; and pronounced jaw movements [63, 100]. The increased intelligibility in Lombard speech is not only linked to acoustic and phonetic modifications in the auditory signal but also results from the articulatory change in the visual speech signal [63, 90, 164]. This thesis will examine and characterise the articulatory modifications in visual Lombard speech; doing so will create a better understanding of visual speech enhancement cues for the support of human perception, since these articulatory modifications are associated with Lombard speech intelligibility. Among other speech examples that exhibit enhancement in the visual domain such as clear speech [128], Lombard speech was chosen in particular in this thesis due the ability to induce such speech under controlled and rigorous experimental settings, given its reflexive nature. One primary issue in the study of visual enhancement in Lombard speech is the current lack of audiovisual Lombard speech data that has been collected in a controlled setting. This lack of data has provided the motivation to collect an audiovisual Lombard speech dataset with plain (non-Lombard) references to each sentence to allow precise characterisation of the speech enhancement made under the Lombard reflex. The collection was made with careful consideration of the various communication factors that could mediate the quality of the produced Lombard speech as well as the saliency of the visual signal. A bespoke head mounted camera system is used to collect both front and profile views of the talkers.

1.2 Contributions

The main contributions of this thesis include the use of two enhancement methods: an appearance based and a kinematics based enhancement method. The effects of these enhancement methods on supporting auditory and audiovisual perception are evaluated using an audiovisual training framework, a training framework which is based on Bernstein *et al.*'s [28] methodology. Another contribution is the bi-view audiovisual Lombard Grid corpus; this in turn serves the final contribution, the analysis of visual Lombard speech. The following sections highlight each contribution and the resulting publications.

Appearance Based Enhancement

The first enhancement is an appearance-based enhancement method that modifies the appearance of the talker's mouth, since the mouth provides a significant proportion of the overall visual speech information gathered from the face [210]. The aim of such an enhancement is to increase the saliency of the visual speech signal. This is achieved by simulating the talker wearing lipstick. An experimental study conducted by Lander and Capek [173] on talkers who wore real lipstick found that the use of lipstick supports lip-reading. The talkers in that study, however, may have been influenced by certain physiological factors that resulted from wearing lipstick, which could have regulated their speech production.

In this thesis, the lipstick effect is applied automatically to the talker, which allows for precise comparison between the intelligibility of speech with and without the lipstick effect. The evaluation of this effect is conducted using the audiovisual training framework designed in this thesis to support the evaluation of the effect of visual speech enhancement in support of the audiovisual perception of CI-simulated speech during training; the post-training effect on improving the auditory perception of CI-simulated speech is also examined. The study was published as:

- Alghamdi, N., Maddock, S., Brown, G.J. and Barker, J. 'Investigating the impact of artificial enhancement of lip visibility on the intelligibility of spectrally-distorted speech'. 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP), 11–13 September, 2015, Vienna, Austria.

The study was also presented as a poster at:

- UK Speech Conference, University of East Anglia, 2015.

A further study on the use of the audiovisual training framework to evaluate the performance of non-native listeners compared to native listeners was published as:

- Alghamdi, N., Maddock, S., Brown, G.J. and Barker, J. 'A comparison of audiovisual and auditory-only training on the perception of spectrally-distorted speech'. International Congress of Phonetic Sciences (ICPhS), 10–14 August 2015, Glasgow, UK.

Kinematic Based Enhancement

The second enhancement is a kinematic-based enhancement approach that exaggerates the speaking style of a talker. This is achieved by following (with modifications) Theobald *et al.*'s method [309], which tested the effect of automatic exaggeration of mouth shapes in videos on the visual perception of lip-readers. In the current study, the audiovisual training framework is used to structure the evaluation of the exaggeration effect in support of audiovisual perception during training as well as auditory perception after the training. The subjects' ability to adapt to the conflict between the articulation energy in the visual signals and the vocal effort in the acoustic signals (because the acoustic signals remained unexaggerated) is also investigated.

The study was published as:

- Alghamdi, N., Maddock, S., Barker, J. and Brown, G.J., 2017. 'The impact of automatic exaggeration of the visual articulatory features of a talker on the intelligibility of spectrally distorted speech'. *Speech Communication* (available online, 13 Aug 2017).

Posters on this study were presented at:

- L'Oréal–UNESCO For Women in Science: PhD poster competition, 2016.
- UK Speech Conference at the University of Sheffield, 2016.

The Bi-view Audiovisual Lombard Grid Corpus

An audiovisual Lombard dataset has been collected in order to investigate real-life examples of visual speech enhancement. Lombard speech was selected in particular among other examples of naturally enhanced speech, as such speech can be induced under a controlled setting. To facilitate a precise analysis, the talkers' head poses in this dataset were controlled, which was achieved by using head-mounted cameras fitted to a helmet designed for this purpose. The dataset includes 55 talkers who uttered 8,250 utterances (4,125 Lombard and 4,125 plain utterances). It offers two views of the talkers (front and side) to facilitate future analysis of speech from different angles. This dataset is an extension of the highly cited audiovisual Grid corpus [54] as it follows the same sentence format as that corpus, although the sentence sets used in this dataset are unique and have not been utilised by the Grid. The plan is to make the audiovisual Lombard Grid dataset available for other researchers.

Visual Lombard Speech Analysis

This thesis characterises visual Lombard phoneme behaviour in different contexts, both within and across talkers. Plain and Lombard utterances collected from eight talkers in the audiovisual Lombard grid dataset (see previous section) were selected for this analysis. A data-visualisation tool was developed to facilitate the extraction of phonemes and word contexts for the talkers' data. The thesis presents a study of visual Lombard speech by considering a number of accounts that provide explanations of visual phoneme behaviour, such as the Hyper-Hypo speech (H&H) theory [187].

1.3 Thesis Structure

The remainder of this thesis is presented in Chapters 2 to 8. The content of these chapters can be summarised as follows:

- **Chapter 2: Speech Perception and Production**. This chapter uses the notion of a 'speech chain' to present the processes that underlie speech perception and production from anatomical and behavioural perspectives; it also presents examples of speech perception chains when undergoing adverse conditions. Doing so will highlight the perception models of CI users and non-native listeners and the characteristics of Lombard speech.

- **Chapter 3: Auditory Training**. A review of auditory and audiovisual training is presented in this chapter. The key factor in training, perceptual learning, is also presented in this review.
- **Chapter 4: Visual Speech Enhancement**. This chapter presents the framework for visual speech enhancement used in this thesis. The chapter presents a review of appearance-based and kinematic-based enhancement methods as well as the design of the audiovisual training framework. Chapter 4 also demonstrates an experimental evaluation of the training framework, which compares the performance of non-native listeners with native listeners.
- **Chapter 5: Appearance Based Enhancement**. This chapter presents the appearance-based enhancement known as the lipstick effect. The chapter starts with a review of facial landmark extraction tools, followed by an elaboration on the technical implementation of the lipstick effect. The experimental evaluation of the lipstick effect using the audiovisual training framework is then presented.
- **Chapter 6: Kinematic Based Enhancement**. This chapter presents the kinematic-based enhancement: the exaggeration effect. The chapter starts with a description of the technical implementation of the exaggeration effect, followed by an experimental evaluation of the effect using the audiovisual training framework.
- **Chapter 7: Visual Lombard Speech**. First, the chapter illustrates the collection of a bi-model audiovisual Lombard grid dataset, covering the equipment used, the collection procedure, and the post-processing of the collected data. An analysis of the visual Lombard speech is then presented by illustrating the selection of the data and the analysis methodology, followed by a presentation of the results and a discussion.
- **Chapter 8: Conclusions**. The final chapter presents the conclusions of the thesis and highlights possible directions for future work.

Chapter 2

Speech Perception and Production

2.1 The Speech Chain

A speech chain, introduced by Dense and Pinson [72], is a linear feed-forward system that describes the processes of speech perception and production [112]. These processes, illustrated in Figure 2.1, are initiated by the talker's thoughts, which are converted into a linguistic format and *articulated* by the vocal tracts' resonances, in conjunction with the external articulators' movements, to produce an acoustic signal [112]. The talker can adapt the speech production given auditory feedback from the produced acoustic signals. The listener receives the acoustic signal by the process of *hearing*, which involves brain activities associated with perception that convert the acoustic signal into a linguistic format and then into meaning [112]. Gick *et al.* [112] revised the speech chain to include one component that has a significant influence on speech perception and production: multi-modality. Speech is intrinsically multi-modal, in which more than one sense contributes in speech production and perception. The revised speech chain is presented in Figure 2.2. Adaptation to the speech chain can also occur under adverse conditions which could affect the clarity of the communication message. Listeners may follow different techniques to recover

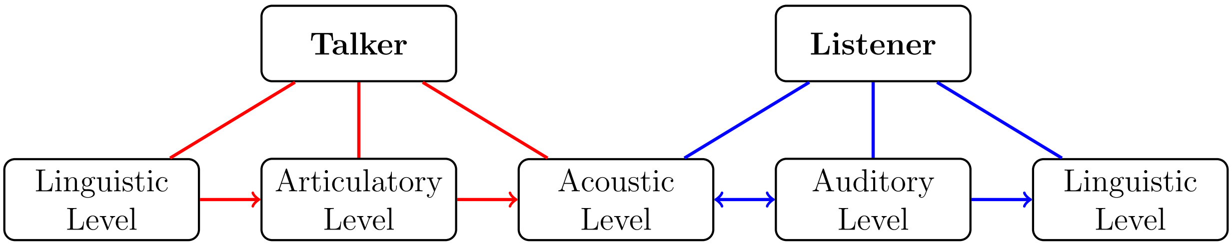


Figure 2.1: The stages of the speech chain.

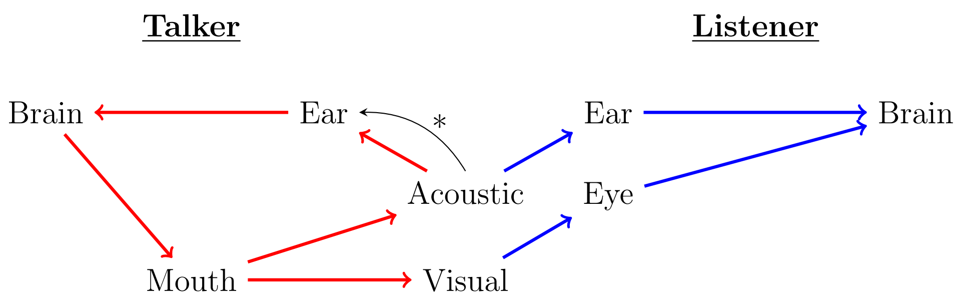


Figure 2.2: Multi-model speech chain with feedback loops, where the * indicates auditory feedback. (Adapted from Gick *et al.* [112].)

the degraded message, while talkers adapt their production in order to increase the intelligibility of their speech.

This chapter presents a review based on the revised speech chain model under normal and adverse conditions. This model is used in particular as it illustrates the interaction between perception and production of audiovisual speech, the communication setting of interest in this thesis. The chain is broken down into the production chain and the perception chain, and key points in both chains that are relevant to the scope of the thesis are addressed. Section 2.2 reviews the anatomy of the brain regions associated with the speech chain, with an emphasis on the role of visual speech in auditory perception. Section 2.3 covers the speech production system and the characteristics of speech; Section 2.4 reviews the speech perception system and the role of visual speech cues in enhancing auditory perception (Section 2.4.1). Section 2.5 addresses possible sources of adversity in the speech perception chain. Two examples of speech perception chain model under adverse condition are selected for this review: the cochlear implant user's perception model (Section 2.5.1) and the non-native listener's perception model (Section 2.5.2). The aim of this review is to highlight evidence from the literature that suggests similarities in audiovisual perception (Section 2.5.3) in these perception models. Section 2.5.4 provides an example of a speech production chain model that acts to counter adversity in perception and offers a real-life example of visual speech enhancement: Lombard speech.

2.2 Language Areas in the Brain

The language zone is the area in the brain that is associated with speech production and perception. Speech perception and production areas in the brain are therefore

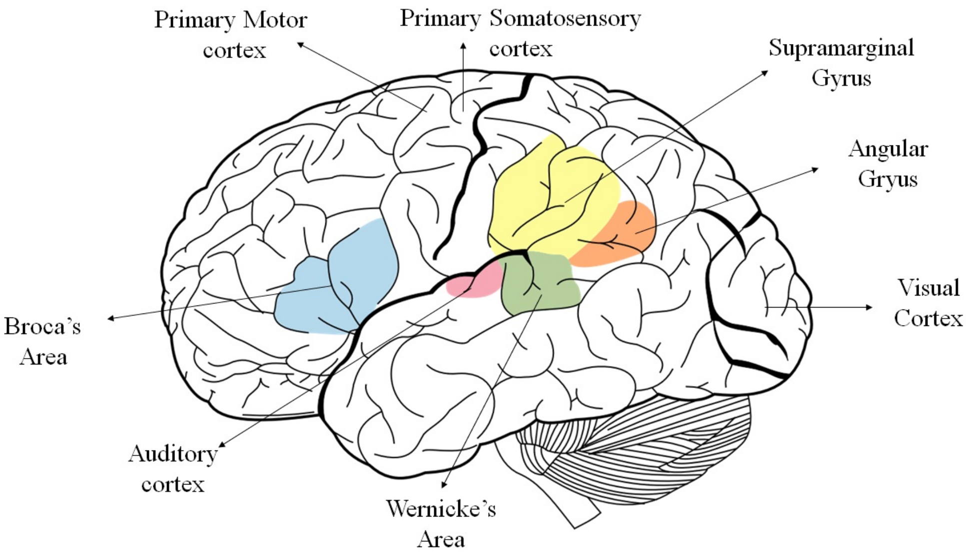


Figure 2.3: Speech and language areas in the brain. (Source: Wiki Commons- released to the public domain.)

interconnected, and thus presented together in this section. The speech and language function in the brain is associated with the *Perisylvian* zone which is the area of *Sylvian fissure* that includes *the auditory cortex*, *Wernicke' s area*, *Broca' s area*, the *Supramarginal Gyrus* and the *Angular Gyrus* (Figure 2.3)[1] [112]. Other areas in the brain associated with language include *the visual cortex*, *the primary somatosensory cortex* and *the primary motor cortex* [112]. The following briefly describes the function of each part according to Gick *et al.* [112]:

- The auditory cortex processes acoustic information and performs basic and high-level functions of audition [248]. The auditory cortex also responds to somatosensory information including facial visual cues [46, 95, 274], suggesting that visual speech cues might be fed to early stages of acoustic speech processing [29].

- Wernicke' s area and Broca' s area are located in the posterior and inferior parts of the left temporal lobe and are connected to each other via a nerve fibre called the *arcuate fasciculus*. Wernicke' s area is responsible for conscious speech comprehension, while Broca' s area is specialises in conscious speech production [112].

[1] By James.mcd.nz [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY-SA 4.0-3.0-2.5-2.0-1.0 (http://creativecommons.org/licenses/by-sa/4.0-3.0-2.5-2.0-1.0)], via Wikimedia Commons. Text on the image is added by the thesis author.

- The Supramarginal Gyrus and the Angular Gyrus are located in the partial lobe. Both areas are responsible for processing high-level information of speech such as phonological processing and emotional responses. The Angular Gyrus also plays a role in multi-model integration of speech [112].

- The visual cortex, located in the occipital lobe at the posterior of the brain, is responsible for processing visual speech information, and shows a stronger response to visual speech when the acoustic speech is compromised by noise [277].

- The primary somatosensory cortex, located at the partial lobe, plays a role in the processing of tactile information during speech perception and in the feedback system in speech production. This cortex shows a strong response when audiovisual speech integration fails [29].

- The primary motor cortex, located parallel to the primary somatosensory cortex, is responsible for sending the speech production plan processed by the Broca' s area to the lower parts of the brain and then to the body limbs associated with speech production [112].

2.3 The Speech Production Chain

The organs involved in speech production include the brain and the associated parts of the nervous system, the respiratory system (diaphragm, lungs, ribcage and trachea), the larynx, and the pharynx (laryngeal, nasal and oral parts) [112]. Figure 2.4 illustrates parts of this system. Speech production involves four processes: respiration, phonation, resonance and articulation. The following briefly explains each process according to Williams [330] and Fernando [314]:

- In respiration, the air is exhaled by the lungs. The manner of respiration is language dependent, for example, English speech sounds result from a 'pulmonic egressive air stream' (i.e., outward-flowing air-stream) while in other languages, such as Scandinavian languages, speech sounds are formed by ingressive sound [113].

- In phonation, the air pressure from the lungs through the trachea is modulated by the closing and the opening of the vocal folds at the larynx. The state of the glottis (i.e., the gap separating the vocal folds) can regulate the frequency of the folds' vibrations and hence the *voicing* of the produced sound: voiced sounds

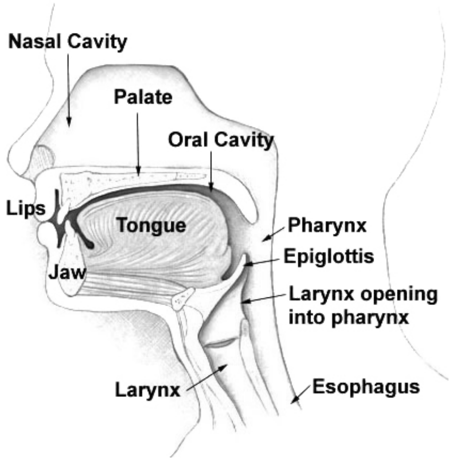


Figure 2.4: Part of the speech production system. (Source: Arcadian [Public domain], via Wikimedia Commons).

are produced when the glottis is slightly opened (increased vibration); voiceless sounds are produced when the the glottis is widely opened (reduced vibration). Other acoustic features that are associated with the vocal cords' movements are the *Fundamental Frequency F0* – defined as the vibration rate of the folds, *Intensity* (or loudness) – defined as the energy of the folds movement, and the *Quality*, which is associated with the movement patterns of the folds.

- In resonance, some acoustic properties of the output sound from the phonation process are further improved by the pharynx system, in particular by the nasopharynx and oropharynx. F0 that is produced by the vibration of the vocal folds is resonated in the vocal tract. The resonance of the vocal tract produces the *formants* (harmonic frequencies).

- In articulation, speech sounds become more distinguished; the configuration of the articulators (lips, teeth, and tongue) can define the manner and the place of articulation.

2.3.1 Speech Segments

The *phoneme* is the basic unit in speech, and can be one of the following categories [254, 263] (phoneme representation here follows Arpabet [168], a phonetic transcription scheme used by the speech recognition CMU dictionary):

- *Vowel*: a sound that features an open larynx and a clear exit to the air pressure with no obstruction imposed by the tongue or teeth. A vowel can be a basic

# MPEG-4	Phonemes (Arpabet)
1	/P/, /B/, /M/
2	/F/, /V/
3	/DH/, /TH/
4	/D/ , /T/
5	/G/, /HH/, /K/, /W/
6	/CH/, /JH/, /SH/, /ZH/
7	/S/, /Z/
8	/L/, /N/, /NG/
9	/R/, /Y/
10	/AA/, /AE/, /AH/, /AO/, /AY/
11	/EH/, /ER/, /EY/
12	/IH/, /IY/
13	/OW/, /OY/
14	/AW/, /UH/, /UW/

Table 2.1: Phoneme to viseme mapping. (The table is adapted from Deena [68]).

sound (*monophthongs*) such as /AA/ and /UW/, or a composite of two vowels (*diphthongs*) such as /AY/ and /OW/.

- *Consonant*: a sound that features a closed or semi-closed larynx. A consonant can be *nasal*, such as /M/ and /N/, when the velum directs the produced air towards the nasal passage; or *fricative*, such as /F/ and /S/, when the produced air flow passes through a constricted exit to produce a friction sound; or *affricative*, such as /CH/ and /JH/, when the produced air flow is constricted then released to produce friction sound; or *plosive*, such as /B/ and /D/, when the airflow is completely blocked then released, which creates an explosive sound.

- *Semi-vowel*, or a 'vowel-like' consonant: a sound that shares the phonetic nature of the vowels but appears within word levels at consonant positions. Examples of semi-vowels are /W/and /Y/.

In connected discourse, the brain organises speech sounds (consonants and vowels) into streams, or speech units, such as syllables. During this process, the articulation of a corresponding phoneme is influenced by the adjacent phonemes. This phenomenon is called *co-articulation*. Co-articulation can be either backward or forward, depending on the position of the influencer phoneme neighbour. Backward co-articulation when the influencer phoneme occurs before the target phoneme, and forward is when it occurs after the target phoneme [124]. This phenomena suggests that speech has a dynamic nature rather than being static: a plan for each speech segment is made even before it occurs [330].

Given a phoneme signal that spans over a number of visual frames (video or animation frames), a *viseme* [89] is a unit of visual speech that describes the articulatory configuration in a frame of that phoneme. The notion of viseme was proposed by Fisher [89] who clustered the visually perceived consonants into viseme categories by grouping confusions in the listeners' responses. In computer animation, the use of the phoneme-to-visime mapping is one technique to produced animated visual speech, where each animation key-pose is associated with a viseme.

Modelling the effect of the co-articulation has received a considerable attention in the literature [43, 53, 87, 180, 245]. One example is the Cohen-Massaro model [53] that uses dominance functions defined for each articulator. Each dominance function simulates the impact of the corresponding viseme on speech production. To define the final shape of the mouth, dominance functions are blended by computing their weighted sum. This generates a curve that represents the final speech trajectory [53, 73].

There are 14 categories of viseme defined by the MPEG-4 standard [119, 231] (Table 2.1). In these categories, there is no one-to-one correspondence between phoneme and visemes. This means that each viseme can be associated with more than one phoneme. There is also no consideration to visual co-articulation (the influence of the surrounding visemes on the mouth shape of the current viseme). Another issue in a phoneme-to-viseme mapping is the natural asynchrony between the auditory and the visual speech signals (i.e., the onset of the mouth movement and the onset of the acoustic production of speech are not aligned). The use of dynamic visemes [307] was proposed to solve the phoneme-to-viseme mapping issues by modelling co-articulation and audiovisual asynchrony using a data-driven method.

2.4 The Speech Perception Chain

Auditory speech perception is carried out by the auditory system, which is composed of the ears and the auditory parts of the sensory system. The human ear (Figure 2.5[2]) consists of the external ear, the middle ear, and the inner ear. The following briefly illustrates the main parts of this system:

[2]By Iain at English Wikipedia [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/)], via Wikimedia Commons

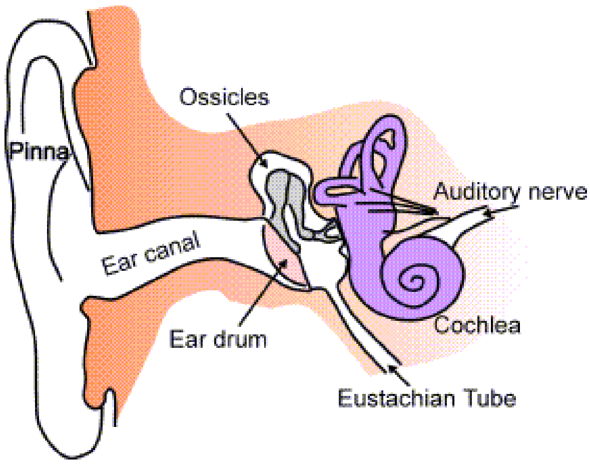


Figure 2.5: Anatomy of the ear. (Source: Wiki Commons – released to the public domain.)

- The external ear is the visible part, which includes the pinna, the external auditory canal and the outer layer of the eardrum. The pinna acts as a funnel, which collects, amplifies and then directs sounds to the ear canal that connects the pinna with the eardrum in the middle ear. The pinna plays an important role in sound localisation by adding directional cues for the perceived sounds [81, 248].

- The middle ear spans from the eardrum to the oval window of the cochlea. It contains the eardrum and the three ossicles (small bones) that are responsible for converting the vibration of the eardrum when sounds are perceived into an amplified pressure energy [81].

- The inner ear comprises the cochlea and the vestibular system, performing the functions of sound detection and balance [311]. The cochlea has a spiral shape: its base is located near the oval window and the other end of the spiral is called the apex. The basilar membrane in the cochlea vibrates in response to the pressure energy that is transmitted to the cochlea fluid from the ossicles. A topographical mapping[3] of frequency (or frequency-to-place) is applied to the basilar membrane surface, starting with high frequencies at the base and graduating to low frequencies at the apex (Figure 2.6[4]). Hence, when the pressure energy arrives, it will travel from the base until it reaches the region

[3]Topographical mapping refers to the spatial organisation of different frequencies processing points in the brain.

[4]By By Kern A, Heid C, Steeb W-H, Stoop N, Stoop R [CC BY 2.5 (http://creativecommons.org/licenses/by/2.5)], via Wikimedia Commons.

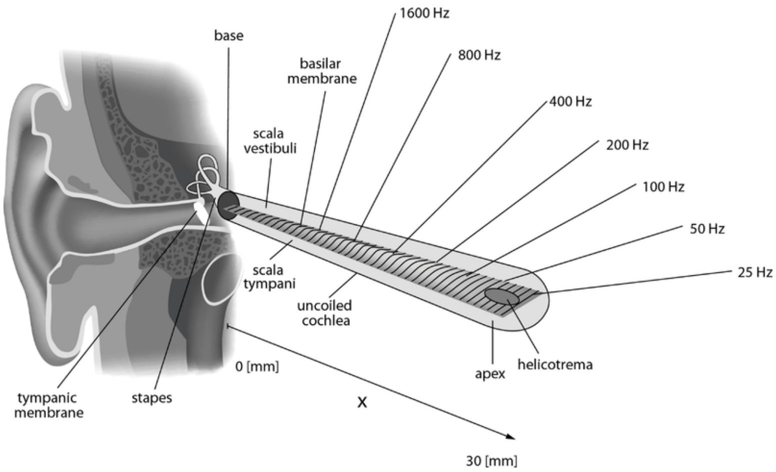


Figure 2.6: Uncoiled cochlea with basilar membrane showing frequency regions. (Source: Wiki Commons – released to the public domain.)

for the corresponding frequency of the perceived sound. Sound transduction is performed by hair cells (sensory cells) that cover the basilar membrane. They respond to the movement of the cochlea fluid by increasing/decreasing the firing rate to the auditory nerve to send sound information to the brain [248].

- The encoded sound information travels from the cochlea to the Central Auditory System (CAS). The CAS represents the auditory pathway from the cochlea to the auditory cortex which is responsible for processing auditory information (Section 2.2). The main functions of the CAS are layered and include sound localisation, pitch processing and multisensory information integration [248].

2.4.1 Audiovisual Speech Perception

In face-to-face communication, speech perception is carried out by both ear and eye (Figure 2.1). Audiovisual (AV) integration refers to a listener' s ability to utilise available sensory information to interpret the perceived message from the talker [172, 204, 267]. The development of AV integration starts in early childhood; Schorr *et al.* [278] found that children develop this ability from birth to 2.5 years of age. This ability is then subject to adaptation and attenuation by the perceiver during communication in adverse listening situations [206]. The importance of visual speech becomes apparent in difficult listening situations (Section 2.5), however, this contribution is also reported during the perception of perfect audible signals. *The McGurk effect* [211] is vital evidence of multi-sensory cortical processing that occurs during speech perception [28]: it is the illusion of perceiving a new audio stimulus when in-congruent audiovisual stimuli are presented. Listeners reported receiving 'ada' when audio 'aba' visual 'aga' were presented, and 'ata' when audio 'apa' visual 'aka' were presented. The McGurk effect suggests that the visual speech signal is not just an addition to the auditory signal; they both complement each other [265, 299]. Indeed, cortical areas in the brain associated with speech show responses to both visual and auditory speech signals [46, 95, 274] (see Section 2.2). Subsequent studies to McGurk and MacDonalds' [211] provide accounts for audiovisual speech integration [181, 186, 202, 205, 244, 299]. For example, Massaro [202] proposed the *Fuzzy Logical Model of Perception* (FLMP) in which acoustic and visual speech features are evaluated by listeners before being integrated.

The role of the visual speech signal in enhancing speech perception is sourced from the strong correlation between the auditory and the visual signals. Visual information extracted from a talking mouth is found to correspond to the temporal envelope of the speech signal (the plotted visual and acoustic data make compatible shapes) [77, 244]. Also, visual speech information can enhance phoneme recognition: Summereld [299] hypothesised that acoustic and visual speech information are complementary in speech perception in which visual cues inform the place of articulation and the acoustic cues inform the manner of articulation [244]. Section 2.5.3 presents further review of audiovisual speech processing.

2.5 The Speech Chain Under Adverse Conditions

Adaptation to the speech chain occurs when the produced acoustic speech signal becomes unintelligible due to adverse conditions. According to Mattays *et al.* [208], adversity can be external or internal to the listener. External adversity originates from source (talker) related factors such as when the talker produces accented, disfluent, or impaired speech. It can also be due to environmental factors that reduce the intelligibility of the acoustic speech signal such as energetic or informational masking. Internal adverse conditions are due to listener-related factors such as when the listener experiences sensorineural hearing impairments, reduced non-native linguistic knowledge, or cognitive load.

This section focuses on the effect of adverse conditions on the speech chain. The perception models of CI users with sensorineural hearing impairments and non-native listeners with reduced non-native linguistic knowledge is reviewed. These two models are of interest in this thesis as they share similar characteristics, including:

- *An impact on perception*: According to Mattays *et al.* [208] both adversities cause:
    - Failure in speech recognition due to a failed mapping between the low level acoustic and phonetic cues to the high-level representation of speech; and
    - Reduced memory capacity;
- *The change in behaviour under external adverse conditions:* CI users and non-native listeners' perception suffers increased deterioration when experiencing external adverse conditions compared with native listeners; and
- *Audiovisual speech perception benefit:* both seem to benefit from the introduction of visual speech cues.

This section also focuses on a speech production model example that demonstrates phonetic, acoustic and articulatory enhancement to counter external adverse conditions such as background noise. This example of speech is produced under the *Lombard effect*. It is driven by an unconscious reaction to noise, and by the need to maintain intelligible communication in adverse conditions. The following sections review each model separately.

2.5.1 Cochlear Implant Users

In the case of sensorineural hearing loss, some or all hair cells that stimulate the auditory nerve are non-functional. A cochlear implant is a surgically implanted prosthesis that stimulates the auditory nerve by firing electrical pulses, performing the function of the damaged hair cells [185]. The main components of a CI are (Figure 2.7):

- The internal part, which consists of a receiver/stimulator, and an array of between 12-22 electrodes that are implanted next to the basilar membrane in the cochlear to stimulate the hearing nerve. The placement of the electrodes corresponds to a topographical mapping where each electrode covers a band of frequency [185].
- The external components, which consist of a speech processor and a radio frequency coil with a magnet. The magnet joins and aligns the external coil

[4] By BruceBlaus (Own work) [CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons.

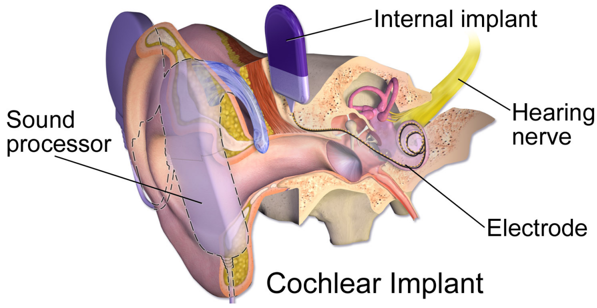


Figure 2.7: Cochlear implant (CI). Source: Wiki Commons – released to the public domain.

with the coil in the receiver/stimulator package. The external speech processor also includes a microphone and other front-end signal processing parts [82].

The stimulation process used by CI processors has a significant role in generating sounds. The following summarises the categories of the stimulation processes used by different makes of CI, based on a review by Choi and Lee [52]:

1. *Stimulating by using a fixed number of channels.* Examples of such processes include the *Continuous Interleaved Sampling* (CIS) method [331] and the *Advanced Combinational Encoder* (ACE) [158]. Both CIS and ACE use a fixed number of channels (16-22) to generate sound by filtering the perceived sound into a number of frequency bands equal to the number of implanted electrodes and map frequencies that are important for speech to each electrode. In ACE, eight to ten frequency bands with the largest amplitudes are selected to be stimulated, whereas in CIS all frequency bands are stimulated.

2. *Stimulating strategy using virtual channel.* The process in this category uses the *virtual channel technique* [76] in which an intermediate virtual channel is created between two active electrodes to compensate for the missing frequencies between those electrodes. An example of such a strategy is the HiRes120 [169].

3. *Hybrid stimulating strategy.* The process in this category uses the virtual channel technique with more than two adjacent electrodes in order to direct the stimulating current to the target region in the auditory nerve. An example of such a strategy is the Four-Electrode Current Steering Scheme (FECSS).

Despite the effort to compensate for the limited number of electrodes, low spectral and temporal resolution is still an outstanding issue for a CI. As a consequence, the perception of the CI user deteriorates when experiencing background noise and when listening to music or to lexical tone [185]. For example, CI users show a higher speech reception threshold (SRT[5]) in noise by (10-25 dB) compared with normal hearing listeners [290]. The loss of the important spectral cues is driven by the speech processing method used by the CI processor; extracting the envelope and discarding the fine structure that informs the harmonics can negatively affect the perceived pitch content [225, 226]. As a result, the frequency region that an electrode stimulates in the basilar membrane does not always match the frequency content received from the sound processor. An overlap between electrode stimulation areas might also occur, making discrimination between different sounds a challenging process for CI users [136, 185]. Temporal cues responsible for informing pitch are also difficult to acquire. Although temporal cues can be informed by pulse rates fired by the electrodes, the pulse rates are always fixed and too high to be utilised for gleaning pitch information. Alternatively, temporal cues could be derived from the envelope of the pulses, however, they appear to be very weak and only noticeable in pulse trains that originate from apex electrodes, the region of low-frequency sounds [136, 185].

The amount of acoustic information CI users can receive is a function of various physiological, neurobiological and neurocognitive factors, including the age at implantation, the severity of damage to the hearing nerve, the degree of the neuroplasticity change pre- and post-implantation, residual hearing, and the number of implanted and activated electrodes [237, 246]. This has resulted in a significant variability amongst CI users [281]. Such variability constitutes an internal adversity in CI users, which hinders the full utilisation of a CI's benefit [208]. Given that, having a homogeneous CI user group is a key obstacle facing CI researchers. An alternative route is to use CI simulation that performs the function of the CI front-end processing, and normal hearing listeners as subjects. Shannon *et al.* [282] and Drorma and Loizou [78] reduced the spectral information of speech by extracting the temporal envelopes from different frequency bands and used them to modulate either noises of the same bandwidth (*noise vocoder*) or sine waves selected from the centre of each band (*sine-wave vocoder*). The simulated speech is then equal to the sum of the modulated noise or sine waves. Dorman and Loizou [78] found no difference in the intelligibility of two groups of sentences: the first distorted using a noise vocoder and the second using a sine-waver. The sine wave vocoder, however, may better simulate the CI processing compared with the noise-vocoder since the fluctuation of noise in the noise vocoder is not present in a real CI [24, 289]. CI simulation, however, does not necessarily reflect the hearing experience of an actual CI user, whose hearing

[5]SRT is the SNR at which 50% of the spoken words are intelligible.

may be worse than the simulation, due to the internal adversity that CI users may experience [65].

After the implantation, CI users undergo *Auditory training*. This is rehabilitative therapy that aims to optimise the listening experience of acoustic signals perceived by the CI user. It is worth mentioning that the stimulation provided by a CI contributes to CAS plasticity [313]. It modifies the physiological response to auditory stimuli and enhances the perception. Auditory training has a similar impact too [313]. Chapter 3 will present a review of auditory training and its impact on auditory system plasticity.

2.5.2 Non-native Listeners

A person is considered a non-native talker/listener of a language when s/he is 'not having spoken that language from early childhood' [238]. Speech produced in a non-native language is known to be less intelligible than speech produced in the listener's native language [93, 294, 319]. Moreover, the intelligibility of non-native speech is known to deteriorate in background noise [106, 175, 209, 319], even for bilingual listeners [209, 261]. For example, Lane [175] found that word recognition of accented speech dropped by 50% in -20 dB SNR; Gat *et al.* [106] found a significant deterioration in non-native word discrimination at 0 dB SNR by non-native listeners who showed comparable performance to native listeners in baseline conditions; Wijngaarden *et al.* [319] found that non-native listeners need better SNR quantified by 1-7 dB in order to achieve 50% in sentence recognition compared with native listeners.

Mattys *et al.* [208] surveyed different accounts that provide an explanation for the effect of 'non-nativeness' in speech perception. First, reduced linguistic knowledge in non-native listeners may contribute to speech recognition failure; a high potential for mapping failure between acoustic/phonetic features and high level representation might occur, as acoustic and phonetic features might deviate from the non-native listeners expectations. Non-native listeners also seem to be more sensitive to distractors, such as competing talkers, which in turn can reduce their attention capacity [56, 178]. They also experience an additive memory load [94] emerging from the normalisation process of linguistic cues across talkers[6] [239] and the complexity of the perception task [177].

There are many factors that can regulate the effect of non-nativeness in speech perception. These include the age at when the non-native language is acquired

[6] Talkers normalisation is a process initiated by listeners to identify words spoken by different talkers despite the acoustic variation across talkers [149].

[209], the relationship between the phonemic inventories in native and non-native languages [319], the experience in the native and non-native language [212], the type of background noise (if present) [56, 177], and the perception task [177].

One important perception adaptation strategy utilised to counter internal adverse conditions is audiovisual speech perception. Under such adversity, listeners assign more weight to the visual speech cues under noisy conditions than in clean conditions, since visual cues remain unaffected by the acoustic adversity. This is proven anatomically as the visual cortex shows a stronger response to visual speech cues when acoustic adversity is experienced [29]. The next section covers audiovisual speech perception under adverse conditions.

2.5.3 Audiovisual Perception in Adverse Conditions

The importance of visual speech becomes apparent in adverse conditions [45, 266]. The pioneering research by Sumby and Pollack [296] showed that the introduction of the talker's face in low SNR conditions significantly improved word recognition by normal hearing listeners from zero to 70–80%. They estimated the effect of introducing the visual signal on speech intelligibly to be the equivalent of increasing the SNR by 15 dB. They also found that the contribution of the visual signals becomes stronger as the SNR levels drop. Subsequent research [36, 85, 198, 210, 221, 224, 266, 297, 298, 324] and more recent research [27, 47, 48, 134, 196, 269, 288] has reported similar findings in normal hearing listeners.

The external articulators (lips, teeth, and tongue) can provide a significant proportion of the overall visual speech information gathered from the face [210, 297, 298]. Summerfield compared the impact of different presentations of visual speech on the recognition scores: full face, mouth only, and points highlighting the centres and the corners of the lips. They found improved accuracy by 43%, 31% and 8%, respectively, indicating that the mouth can only account for 72% of the recognition accuracy in full face mode. McGrath [210] also found that the recognition accuracy of monophthongal vowels reached 56% by lipreading the external articulators only. By visualising these external articulators' kinematic information using a point-light technique, Rosenblum *et al.* [266] found that such visualisation can substantially enhance the intelligibility of speech in noise. Compared with other facial movements that provide temporal cues only, kinematic information from the mouth can provide both speech information and temporal cues [162].

Visual speech cues from external articulators can provide important cues about the place of articulation for consonants, monophthongal vowels, and diphthongs

[48], which can provide significant support when missing phonetic features are compromised by acoustic adversity [134]. This is supported by studies that found a correlation between the physical characteristics of the mouth and the accuracy of lip-reading of vowels [147, 224]. For example, Montgomery *et al.* [224] found a correlation between the accuracy in the identification of the monophthongal vowels and the degree of the horizontal mouth aperture and the rounding of the lips, and the accuracy in the identification of the diphthongs and the degree of the vertical lip aperture and the rounding of the lips.

The perception of CI users substantially improves in face-to-face communication [18, 74, 77, 80, 141, 153, 261, 271, 317] (a recent review is provided in [293]). Also, CI users are better multi-sensory integrators [153, 261] and show a stronger McGurk effect [271, 295] than normal hearing listeners. They are also more biased toward visual information than acoustic information in audiovisual integration [271, 295]. Recent evidence suggests that seeing the talker's face can activate the auditory cortex response in CI users [295]. The introduction of visual speech facilitates CI users' perception of visually distinguished phoneme categories such as anterior consonants like bilabials, and posterior consonants. Some phoneme categories, however, such as /b/,/p/ and /m/, dental, and velar consonants are more challenging for CI users. This is because they require voicing and manner of articulation in order to be identified [74, 291]. The visual speech signal also improves CI users' recognition of syllables and hence enhances lexical segmentation [77].

Individual differences between CI users in utilising visual signals have also been reported [153, 204, 205]. Factors that regulate the benefit of the visual signal for CI users include the duration of deafness and duration of CI usage [74], the onset of deafness (pre- or post lingual) [25], the degree of cross-modal plasticity[7] acquired during deafness [12, 174, 216], and the perceptual and cognitive abilities of CI users [148, 198].

Non-native listeners also benefit from the introduction of a visual signal in the perception of native speech [130–132, 146, 280, 327, 335, 336]. Listeners from different languages showed a stronger McGurk effect when listening to a non-native language than when listening to their own language [66, 130, 280]. There are many factors that can regulate the non-native listener's sensitivity to visual speech cues, including the level of native linguistic abilities (native language proficiency) [125, 240], the correlation between the phonemic inventory in the native and non-native language, the visual saliency of the non-native phonemic contracts, the degree of compatibility

[7]More information about neural-plasticity is provided in Chapter 3.

between visemes in the native and non-native language [131, 132, 327], and the utilisation of visual cues in the native language [130]. In addition to these factors, audiovisual language training can help improve the utilisation of visual cues by non-native listeners as it increases the exposure to the native language and hence increases native language proficiency [131, 132, 327].

2.5.4 Production Under Adversity

According to Lindblom' s hypo- and hyper-articulation (H&H) theory [187] of speech production, speakers make articulatory energy modifications from hypo- to hyper-articulated speech in order to adapt to a listener's commutation needs or to environment conditions [41, 55, 91, 128, 164, 194]. Clear speech is one example of hyper-articulation which results from addressing a listener's needs, such as limited linguistic knowledge (for example, when the listener is an infant or a non-native listener), hearing impairment, or when the listener is situated in external adverse conditions [128]. For example, Hazan and Baker [128] found evidence of acoustic and phonetic modification of talkers' clear speech when listeners experience adverse conditions such as vocoded speech (CI simulated speech) or babble speech, even when the talkers remained unaffected by the adversity [121, 128, 129]. Acoustic and phonetic speech adaptation techniques observed in clear speech include increased F0 and speech level, and decreased speaking rate [35, 55, 128, 193]. Evidence of visual articulatory adaptations in clear speech is also found which involve larger lip and jaw movement and wider inter-lip area [129, 306].

Hyper-articulation is also observed in noise-induced speech, i.e., *Lombard speech* [101, 192, 283]. The Lombard effect, named after Étienne Lombard, is the reflexive adaptation to speech production with the aim of countering reduced speech intelligibility under noisy conditions [41, 192]. The mechanism of Lombard speech is driven by two loops: a private loop in which the Lombard speech is regulated in response to the auditory feedback in the speech chain (Figure 2.1); and a public loop in which Lombard speech is regulated in accordance to the listener's needs [176]. Lombard speech is characterised by a collection of acoustic and phonetic modifications including [17, 55, 62–64, 151, 152, 161, 161, 164, 193, 194, 260, 283, 286]:

- An increase in F0;
- An increase in the signal energy;
- A shift in the first and the second formant (F1 and F2);

- A tilt of the speech spectrum that boosts higher frequencies;
- An increase in vowel duration; and
- Energy shifts amongst different classes of phonemes.

In the visual domain, a body of literature has considered visual articulatory changes in Lombard speech [62–64, 100–103, 121, 140, 161, 163, 164, 283, 284, 320]. For example, by analysing motion data elicited from markers placed on a talker's face, greater face and head motion was observed in visual Lombard speech [320]. In a series of studies on French talkers, a greater global change in the movement of the jaw and the lips were found by analysing the amplitude and the velocity for lip spreading, lip aperture, inter-lip area and lip pinching elicited from recorded videos of the talkers [100–102]. Studies on English-Australian talkers also found increased jaw and lip motion and protrusion after analysing the Principal Components (PC) of motion data acquired from motion sensors placed on the talkers' faces [63, 163, 164]. The degree of visual modification in Lombard speech, however, is not uniform across articulators [320]. For example, jaw movement and lip spreading and opening were found to be greater than lip protrusion [100, 163]. Huber and Chandrasekaran [140] found greater displacement and velocity of the lower lip movement. Simko *et al.* [283, 284] analysed articulatory trajectory data tracked from sensors placed on the lips, jaw and the tongue of Slovak talkers and found that the movements of the lips and jaw were greater than the tongue, a similar finding to Garnier *et al.* [103]. Correlations between acoustic and visual features of Lombard speech have also been reported. For example, a correlation was found between RMS speech intensity and the PCs of jaw and mouth [63], and RMS speech intensity and face and head motion [320].

The Lombard effect has a significant impact in improving the intelligibility of acoustic speech produced in adverse conditions. This is driven by the acoustic and phonetic adaptation induced by the Lombard effect [55, 193, 194, 301]. An increased benefit of visual speech was also reported [91, 92, 164, 321]. Although Vatikiotis *et al.* [321] found no difference between the visual benefit of plain speech and Lombard speech, Kim et al. [161, 162, 164], Fitzpatrick *et al.* [90–92] and Davis *et al.* [62–64] reported an increased benefit of visual Lombard speech in supporting speech intelligibility.

Although studies on mammals have shown that the neuronal circuits responsible for inducing the Lombard effect are situated in the brain stem, indicating that it could be a physiological reflex, it has been found that the Lombard effect can be controlled and regulated [41]. This is evident in studies that report how the Lombard effect on acoustic, phonetic, and articulatory adaptations of speech is a function of communication environment variables, including masker type [62, 63, 193, 194], masker immersion method [63, 102], masker level [223, 284] communication task [63, 102], communication modality [91, 92, 103], and words and utterance segments [101], as well as inter-talker variables such as gender and language [152]. For example, audiovisual speech modifications are found to be more intense under a 'cocktail party' masker than a white noise masker [63], when the masker was presented via headphones compared with loudspeakers [63, 102], and in the last syllable of the target word than other syllables [101]. Audiovisual speech modifications when the talker is involved in a communication task have also been found to be amplified and include more information than audiovisual speech modifications made when the talker is reading sentences [102]. The impact of communication modality remains contentious, as some studies found a greater saliency of visual Lombard speech in face-to-face communications [91, 92], while Garnier et al. [103] suggested that visual modifications are just correlates to the acoustic adaptations that are greater in the audio-only modality. By studying the impact of masker levels, Simko *et al.*[284] found a non-linear effect of masker level on articulatory movement.

A number of enhancement algorithms have sought inspiration from the acoustic, phonetic and articulatory features of Lombard speech to enhance the intelligibility of the acoustic speech [6, 116, 133, 139, 286], and to synthesise acoustic Lombard speech [227, 236, 247]. The only research found that aimed to synthesise visual Lombard speech was done by Alexanderson and Beskow [13]; they made video recordings of a talker uttering short sentences in Lombard and plain conditions, and used facial motion data elicited from that talker to animate a 3D avatar. They tested the intelligibility of two types of animations: type 1 – congruent animation in which visual Lombard animation is coupled with Lombard speech; and type 2 – in-congruent animation in which visual Lombard speech is integrated with plain speech. In the in-congruent animation case, the Lombard videos were time-warped in order to be aligned with the plain audio. The audio part of the animation and the original videos were then acoustically degraded using a noise vocoder to reduce their intelligibility prior to a subjective intelligibility test. The test revealed an increased intelligibility in both animation types that were driven from the Lombard video against the animation driven from normal video. Type 1 animation attained comparable intelligibility to the Lombard video and they were both more intelligible than plain videos. Type 2 animation attained similar intelligibly to plain videos but was less intelligible than the Lombard videos.

The collection of modifications in acoustic and articulatory features of speech triggered by the Lombard effect has major implications in speech processing research, in particular, in automatic audiovisual speech recognition (AVSR) systems. Such systems are usually trained on clean speech datasets such as the Grid [54], however, their performance can deteriorate under Lombard conditions [5]. The main barrier that faces such research and any Lombard-oriented research is the limited access to Lombard data. Despite the existence of a large body of literature addressing the analysis of auditory and visual characteristic of visual speech, data used in such research is not available. Very limited resources for Lombard speech data are available to the public. One example is AVICAR [1, 179], which is an audiovisual speech corpus recorded in a car environment. It features 100 talkers reciting in English 10 isolated digits, 26 isolated letters, 20 phone numbers, and 20 TIMIT [342] sentences under five driving scenarios. Noise conditions in AVICAR, however, are characterised with low SNR conditions (-10 dB to 15 dB), with no clean reference for the utterances recorded in the noisy conditions. Despite the clear importance of this issue, until now there have been no audiovisual Lombard datasets recorded in a controlled setting with consideration of the communication environment variables.

In this thesis, visual Lombard speech will be considered as an example of visual hyper-articulation. It was favoured over clean speech as a convenient case study to study visual hyperarticulation adaptation, because it can be induced in a controllable manner [284]. In Chapter 6, a Lombard-inspired visual exaggeration method for visual plain speech is presented. The collection of a Lombard speech dataset and analysis of that data are addressed in Chapter 7.

2.6 Summary

This chapter presented a review of speech perception and production in plain and adverse conditions. The review started by addressing the relevant language areas in the brain that showed the interconnectivity between perception and production and the multi-sensory nature of speech. This was followed by an overview of the main processes that are involved in the production of speech and how each process is associated with certain acoustic and articulatory characteristics of speech. The main units of speech were also reviewed as well as the dynamic nature of speech represented in the co-articulation phenomena. An overview of the main components of the speech perception system was then presented with a focus on audiovisual speech perception in normal conditions. Adversity in the speech chain was then addressed by reviewing the possible sources of adversity and their impact on perception and production. Examples of adversity in perception that are relevant to this thesis can be found in CI users and non-native listeners. Both make use of audiovisual speech perception in order to improve the intelligibility of the perceived message. Lombard speech was then addressed as an example of speech production adaptation to counter adverse conditions by highlighting the collection of acoustic, phonetic and articulatory adaptations that accompany Lombard speech and the variables that can regulate these adaptations.

Although CIs have revolutionised the treatment of hearing loss, they do not provide an optimal hearing experience for CI users. CI users still need to undergo auditory training to shape their listening abilities after the implantation. This is expected since CIs only recover hearing, but not listening, which is a vital requirement for a successful communication. Auditory training is covered in the next chapter.

Chapter 3

Auditory Training

3.1 Introduction

Hearing aids and implants show promising outcomes in restoring recipients' audition, yet audition is just one step in a series of events that form an adequate communication experience [302]. Figure 3.1 shows the key communication elements, proposed by Kiessling *et al.* [159] and refined by Sweetow and Sabes [304], in which the listening stage is a key step towards successful communication. The listening stage interacts with the comprehension and communication stages and creates a positive influence on the communication process when linguistic and acoustic features are utilised effectively at that stage. In contrast, it has a negative impact if it fails to do so, even when listening is accompanied by good hearing skills [302]. Therefore, it is essential for recipients of hearing aids and implants to undergo *Auditory Training*.

Auditory training is a speech perception training that helps to optimise the listening experience for hearing aid and CI users. Auditory training utilises *Auditory Perceptual Learning* (APL), which enables training subjects to generalise the learning experience they acquire from the training to new auditory/speech stimuli and listening environments after the training. A correlation between APL magnitude and Central Auditory System (CAS) plasticity has been found, which reflects the significant role of auditory training in improving listening skills [97]. Recent evidence suggests that audiovisual speech in auditory training, or *Audiovisual training*, can enhance APL outcomes [28].

This chapter presents a review of perceptual learning, auditory training and its relation to auditory perceptual learning, as well as the impact of introducing visual speech in auditory training (audiovisual training). Audiovisual training is the context chosen to apply visual speech enhancement in this thesis, as it facilitates access to both auditory and audiovisual perception of listeners. Section 3.2 presents a general review

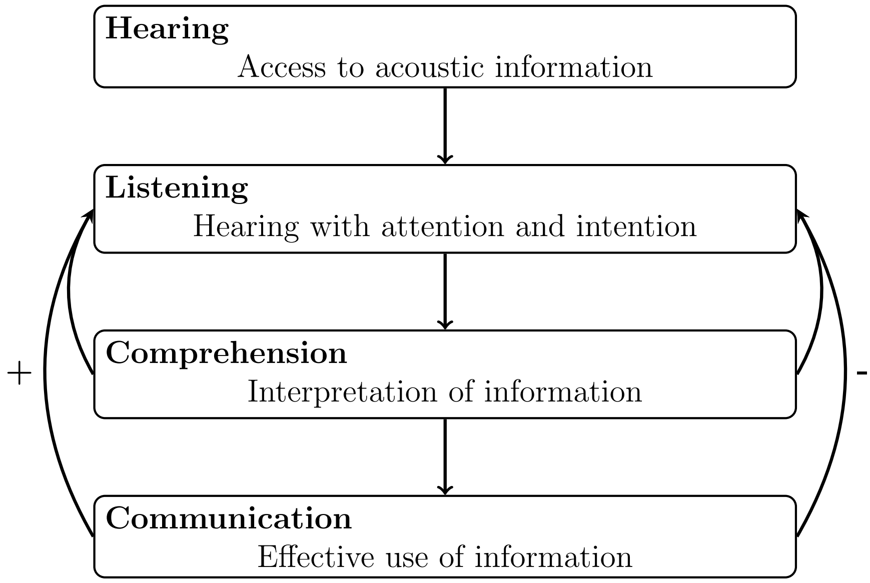


Figure 3.1: Elements of a communication feedback model. (Adapted from Sweetow *et al.* [302].)

of perceptual learning by demonstrating evidence from the literature on its impact on neuroplasticity. Section 3.3 covers auditory training by exploring its processes, and evidence from the literature that describes the benefits of auditory training to CI users. The section covers evidence that reports auditory training impact on inducing the auditory perceptual learning, and the increased effectiveness of the computerised auditory training. Section 3.4 looks at evidence for the increased effectiveness of auditory training when visual speech is presented, i.e., audiovisual training, and also explores works that explain the role of visual speech in guiding effective perceptual learning.

3.2 Perceptual Learning

Perceptual learning contributes to the robustness of speech perception [328]. The theory of perceptual learning [111] refers to 'a practice-induced improvement in the ability to perform specific perceptual tasks' [9]. Perceptual learning aims to enhance sensorial receptors by creating a long-lasting change that optimises the target perception through an intensive and iterative training process [117]. The target outcome of perceptual learning is to generate a perceptual experience that helps the learner to gain sufficient discrimination skills to be able to interpret ambiguous and novel stimuli during the perception process [9, 117].

The mechanism of perceptual learning involves a series of processes, including attentional weighting, stimulus imprinting, differentiation, and unitisation. In attentional weighting, the perception adaptation is induced by increasing the learner's attention towards the important aspects of the perceptual task, and decreasing the attention elsewhere. In stimulus imprinting, perception adaptation can be initiated through the development of receptors associated with certain stimuli. Receptors in this process can be developed for the entire stimulus category, for parts that feature the stimulus, or for the space and position of that stimulus within the global space. In differentiation, stimuli categories that were previously blended become distinguishable and separable. Similar to imprinting, differentiation may occur at stimulus level or feature level. Unlike differentiation, a perceptual task that is intrinsically viewed as a set of parts is viewed in unitisation as a whole unit or category [117].

Perceptual learning is found to induce an interesting phenomena of the human brain: *Neuroplasticity* [44, 49, 114, 184]. Neuroplasticity is the brain's ability to change and to be re-organised by forming new synapses or modifying old ones [218]. Merzenich *et al.* [217] found evidence of enlargement to limbs' neural maps due to extensive limb movement, and shrinking in such maps when these limbs were not used [138]. This suggests a plastic and dynamic nature of the human brain, which promotes alteration and adaptation to neural representation when required. One example of neuroplasticity is in the case of brain injury, in which the brain has the capacity to initiate re-mapping of the neural representation in the damaged part of the cortex in order to compensate for the missing functions. For example, in hearing impaired patients, Central Auditory System (CAS) plasticity has been observed at the event of the deprivation of hearing, and at the acquisition of a hearing aid or cochlear implant [233].

Another event of neuroplasticity is associated with learning and experience [44, 268]. Brain areas associated with certain skills tend to grow as experience in such skills increases. Micgelon's survey [220] provided some examples of this effect. An example is London taxi drivers. MRI shows that they have a larger hippocampus – the brain region responsible for the acquisition of spatial information necessary for navigation – than London bus drivers [200]. Such finding links between spatial knowledge and experience acquired from navigating within a large city and the increased volume of the hippocampus [200]. Bilinguals have also been found to have a larger inferior parietal cortex – the brain region associated with language – than mono-linguals [213]. A difference was also reported in motor, auditory and visual brain areas between musicians and non-musicians [105]. Moreover, neuro-plastic changes detected by EEG signals have been observed following extensive training [11]. For example, changes in event-related brain potentials (ERPs) has been reported following auditory training [312] and in visual evoked-potentials (VEPs) following visual training [253].

Perceptual learning has two phases. The first phase is rapid or fast perceptual learning that occurs within the first hour of training. The second phase is a more gradual and slow learning process that spans several training days, which can result in more improved perceptual skills. Although slow learning is more effective in inducing brain plasticity and enhancing stimuli response [312], evidence has shown that neuroplasticity also occurs during rapid perceptual learning [11, 127].

Evidence provided in this section suggests that perceptual learning can be used as a powerful tool in rehabilitative training. The next section focuses on auditory training and its correlation to perceptual learning.

3.3 Auditory Training

Auditory training is an auditory perception training that helps individuals with degraded hearing abilities to use their residual hearing effectively [42]. Three main patient-therapist practices that constitute a typical auditory training process are:

1. *Listening*: patients are exposed to a repetitive presentation of an adequate number of acoustic stimuli;

2. *Response*: patients provide responses to the presented stimuli; and

3. *Feedback*: patients receive corrective feedback upon their responses from the therapist/therapy provider.

For cochlear implant (CI) users, auditory training is offered during post-implant rehabilitation in order to optimise CI users' speech perception. The required amount of therapeutic intervention is determined according to different variables, including the onset of the hearing loss (i.e., pre or post-lingual hear loss), the age of a CI recipient when receiving the implantation, type of implantation (uni- or bi-lateral), and the residual auditory/lingual abilities [234].

Erber [86] proposed auditory training practices that aim to improve listening skills for hearing aid and CI users. These practices are: sound detection, sound discrimination, sound identification, and sound comprehension. The sound detection phase transfers the training subjects from a situation where they are unaware of the sound to a position of being able to detect it. In the discrimination stage, the training subjects learn to distinguish and discriminate between fine acoustic features, such as temporal and spectral features, which are responsible for differentiating between speech and environmental sounds. Auditory stimuli used in these early stages involve non-speech stimuli such as music and *Ling* sounds (aa, oo, ee, ss, sh, mm) that offer multiple frequency ranges. The training subjects in the discrimination stage are also trained in pre-lexical/phonological levels such as syllables in order to assimilate the blending of speech sounds. At the identification stage, the subjects are trained in the lexical level of auditory stimuli, in which they learn to associate speech stimuli, in closed and open sets, to their meaning. Lastly, the sound comprehension stage trains the subjects to interpret the utterance message. Utterances are introduced in progressively complex levels by different talkers in different environments utilising different conversational cues such as prosodic and contextual cues. [303, 305, 316, 337].

Auditory training practices can be introduced in one of two main approaches: Analytic (bottom-up) or Synthetic (top-down). The analytic approach uses speech elements as the training stimuli with the aim of improving the discrimination skills of the speech, and hence the performance of the peripheral auditory processing. The synthetic approach uses more complex speech stimuli such as sentences with lexical and contextual cues presented in different listening conditions with the aim of improving communication skills such as attention, perception in adverse conditions and use of context, and hence improving the functions of the auditory central processing [270, 302]. The selection of the training approach is dependent on the training goal. For example, the synthetic approach has been widely used in training under adverse conditions [65] while the analytic approach is usually used for training in quiet settings [97].

Evidence from the literature suggests the effectiveness of auditory training in improving speech perception for hearing aid/CI recipients or normal hearing listeners listening to a CI simulation. Improvements in the perception of different speech stimuli reportedly include vowel and constants [97], syllables [333], and words and sentences [65, 135, 291, 303]. Intensive auditory training that involves repetitive exposure to a larger set of training stimuli and immediate feedback can create *Auditory Perceptual Learning* (APL) [98, 233, 312]. Elector-physiological evidence of cortical reorganisation, i.e., neuroplasticity, following auditory training has been found [20, 33, 120, 170, 215, 233, 242, 258, 312]. For example, Menning *et al.* [215] found an increased activity of slow auditory evoked and mismatch field, which are associated with the auditory discrimination process in the brain, following auditory training on frequency discrimination. An increase in the mismatch negativity – an event-related cortical response that correlates to the auditory discrimination – elicited from subjects after undergoing voice onset time (VOT) discrimination training has also been reported [312]. A similar change in the mismatch negativity response was reported by Kraus *et al.* [170] following auditory speech discrimination training, and by Atienza *et al.* [20] following complex auditory patterns discrimination training. These findings suggest the substantial impact of APL resulting from auditory training in inducing cortical plasticity.

A review by Wright and Zhang [334] reported evidence of slow and rapid perceptual auditory learning gained by auditory training. For example, slow perceptual learning of pure tone discrimination [69, 71], Fundamental Frequency (F0) discrimination [71], temporal interval discrimination [155], relative timing task [230] and spatial hearing [334] was reported when subjects were exposed to a number of stimuli ranging between 750 to 1800 stimuli in multi-day training (4-12 days) [334]. Rapid perceptual learning was also reported in pure tone discrimination [70] and spatial hearing [272] after an exposure to a minimum of 200 stimuli in a single day of training. This suggests that APL magnitude can be regulated given the training task and approach [107], the training stimuli [222, 326] and the training frequency [334]. APL outcomes were also found to be generalised and transferred to novel stimuli that were not experienced during the standard training [334]. For example, frequency discrimination learning and stimuli duration learning can be generalised to untrained stimuli [69]. Furthermore, even when learning was conducted at the speech level, it was found that learning generalised to non-speech and environmental sounds [334], which indicates that APL can occur at both phonetic and lexical levels [65, 135, 291].

The use of computerised auditory training is promoted as it provides extensive one-to-one interaction, and a customised protocol in a cost-effective way. A computerised training platform can provide the main auditory perceptual learning principles [303] including the presentation of a large set of stimuli in a repetitive manner, providing immediate feedback for subjects responses, adjusting the task difficulty based on subjects responses, the provision of different listening conditions and speaker variability, tracking patients' performance, and the provision of remote monitoring to therapists [42, 303, 304]. Additional logistics and social benefits include the accessibility of the training in accordance with patient time, and the involvement of patients' significant others in the training, which is a vital factor in rehabilitation success. It is not a surprise that computerised training is now replacing traditional therapy[1]. Examples of computer-based auditory training include LACE [304], Rannan [10, 122] Fu *et al.* [97] and MOGAT [340]. Higher training gain by computer-based training subjects than those attending classical therapy has been reported, which indicates the provision of a more robust auditory perceptual learning by computerised training than with classical training [122, 304].

Another useful application of auditory training has also been reported in improving the perception and the production of speech in a language by learners of that language [132]. Auditory perceptual learning gained from auditory training helped to improve the perception of non-native phonemic contrasts that do not exist in a learners' native language (for example the English /l/-/r/ contrasts for Japanese and Korean learners) [34, 125, 131, 132, 189, 191]. An improvement has also been reported at the speech production level [34, 132]. A greater improvement in the auditory perceptual learning outcomes was found when the associated visual speech (the talker's face) was coupled with the acoustic stimuli and presented in the training [125, 131, 132]. Such learning was also generalised to novel talkers and new stimuli after the training, and to speech articulation by listeners [125, 132]. There are some factors that could regulate the benefit of visual cues for a language learner during the training, including the visual saliency of the non-native phonemic contrasts, the correlation between the phonemic inventory in the native and the non-native languages, and the level of native linguistic abilities [131, 132]. All this evidence leads to an important finding: *Exposure to audiovisual speech can improve auditory-only speech perception*. These findings have encouraged researchers to investigate the impact of using audiovisual speech in rehabilitative auditory training, i.e., *Audiovisual training* (AV). The next section presents AV and its role in improving auditory training outcomes.

3.4 Audiovisual Training

Until recently, the use of AV training, with natural or synthetic 3D audiovisual stimuli, in the rehabilitation domain was restricted to speech-reading and speech training [84, 88, 203]. A popular example is Baldi [203], which is a general-purpose speech/language tutor embodied in a 3D talking head with synthetic and natural speech that has been used in speech training for CI users. A recent line of research has demonstrated another potential rehabilitative application for AV training:

[1] As an example, the cochlear Implant centre in King Saud University Hospital in Saudi Arabia has reduced the use of the clinical auditory training therapy by offering auditory training software to CI recipients [122].

evidence shows a link between AV training and an improved post-training auditory perceptual learning of audio-only CI simulated speech by normal hearing listeners [28, 156, 250, 264, 328]. One of the early studies that reported this effect was done by Rosen *et al.* [264], in which subjects seated inside a booth were asked to listen-to-then-repeat CI simulated connected speech, distorted using a four-channel noise vocoder and produced by a talker sitting outside the booth and facing a subject through a glass window. Pre- and post-auditory-only test results showed a significant improvement in recognition scores (1% to 40%, respectively). The main limitation of this study, however, is the absence of a baseline or control group (AO group) to evaluate the results [250]. With the provision of control groups, subsequent studies [28, 156, 250, 328] (summarised in Table 3.1) featuring different training methodologies, stimuli type, and CI simulation methods, suggest the impact of AV training in boosting the auditory perceptual learning of CI simulated speech. Evidence of generalised learning to novel stimuli that were not presented in the AV training was also reported [156].

Bernstien *et al.* [28] used the Reverse Hierarchy Theory (RHT) of perceptual learning to explain how multi-sensory stimuli, audiovisual in this context, can facilitate uni-sensory perceptual learning, such as auditory perceptual learning. The RHT of perceptual learning states that immediate perception in a sensory pathway (Figure 3.2) depends on high-level information, in which a mapping between the high and low level information of the perception task exists. If not, then perceptual learning is initiated to establish this mapping [21, 28]. In the case of CI simulated speech, higher-level information in the auditory pathway is compromised due to the vocoding process, therefore, an external support is required to guide the perceptual learning. Evidence has shown that a higher-level representation in a sensory perception pathway can be utilised to support the acquisition of a low-level representation in another sensory perception pathway (Figure 3.2) [28, 108, 278]. To achieve this, remapping is done via a backward search initiated by a sensory path that was not affected by the distortion, i.e., the visual path. For example, in the visual pathway, the phonetic cues can provide an access point to the corresponding articulatory features, and consequently to the acoustic cues in the auditory pathway, given the natural correlation between the acoustic and articulatory features. Therefore, a mapping between the auditory phonemic features and acoustic features can now be re-formed and auditory perceptual learning for this task is hence achieved [28].

[2] Bamford-Kowal-Bench (BKB) sentences [23].

Study	Subjects	Training stimuli	CI simulation	Duration (in days)	Method: AO pre test	Method: Training modalities	Method: AO Post test	Gain
Kawase *et al.* [156]	34	500 words	2-band noise vocoder	1	✓	G1. AO w/o feedback; G2. AV w/o feedback; G3. AO w/ feedback; G4. AV w/ feedback.	✓	G1 = 10%; G2 = 35%; G3 = 65%; G4 = 85%.
Pilling *et al.* [250]	42	76 BKB [2] sentences	8-band noise vocoder		✓	G1. natural AO; G2. AO; G3. AV.	✓	G1 = 29%; G2 = 33%; G3 = 62%.
Wayne *et al.* [328]	90	50 simple sentences	4-band noise vocoder.		-	G1. AV w/AO feedback; G2. AV w/o feedback; G3. Natural AO + AO w/ AO feedback; G4. Natural AO + AO w/o feedback; G5. AO w/ (Natural AV + AO) feedback.	✓	G1 = 60%; G2 = 30%; G3 = 40%; G4 = 40%; G5 = 25%.
Bernstein *et al.* [28]	37	72 nonsense words	12-band sinewave vocoder	4	✓	G1. AO; G2. AV. 3 blocks of training followed by AO test.	✓	G1 = 12%; G2 = 19%.

Table 3.1: Summary of Audiovisual training studies. Keywords: Plain: not CI simulated; w/: with; w/o: without.

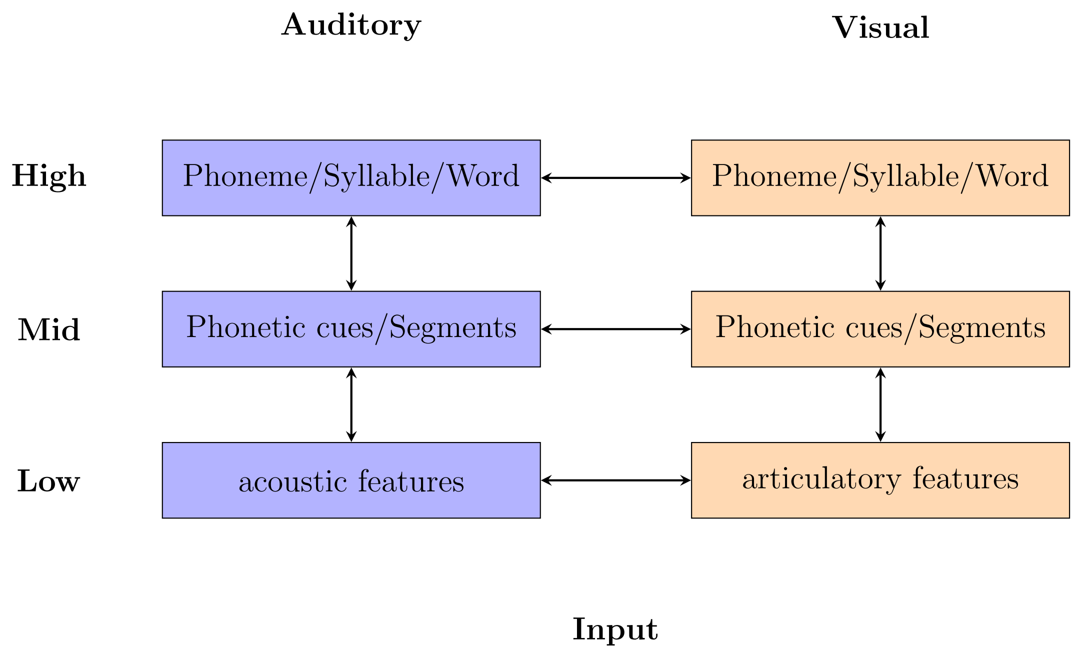


Figure 3.2: A higher-level representation in a sensory perception pathway can be utilised to support the acquisition of a low-level representation in another sensory perception pathway. (Adapted from Bernstein *et al.* [28].)

Many factors may control the AV benefit in enhancing subsequent AO perception of CI simulated speech. For example, Kawase *et al.* [156] found that the provision of feedback contributed to improving the training outcome in both modalities – AO and AV – with a greater enhancement for those doing the AV training. Wayne *et al.* [328] tested several training modalities in which CI simulated stimuli were either presented alone, or after a plain presentation of that stimuli, and with and without the provision of feedback. Consistent with [156], AV training subjects outperformed the other training groups only when feedback was offered in the training. The training methodology can also affect the impact of AV training. Although Bernstein *et al.* [28] found enhanced auditory perceptual learning for subjects trained by AV training (Table 3.1), in a subsequent experiment, they varied the AV training methodology such that alternating AV and AO training blocks were presented, and measured the impact of this variation against AO training. The results suggested that the introduction of AO training blocks within the AV training impeded the development of auditory perceptual learning.

The above evidence suggests the great potential of AV training in improving auditory only-listening skills. As reported in Section 3.3, Hazan *et al.* [131, 132] found that audiovisual training became more beneficial when subjects trained on discriminating visually salient phonemic contrasts (such as /b/,/p/ and /v/) than when trained on less salient contrasts (such as /l/ and /r/). This may suggest that increasing the visual saliency of visual cues in audiovisual training may help to increase its benefit. To the best of our knowledge, enhancing the visual cues in order to increase their saliency in audiovisual training aimed at enhancing post-training audio-only listening skills has not been previously addressed. This has provided the motivation to investigate the effect of visual speech enhancement in the domain of audiovisual training, as will be presented in Chapters 4, 5 and 6.

The perceptual learning phase that will be examined in this thesis is the rapid phase that occurs within the first hour of the training (Section 3.2). Therefore, a shortened version of auditory training conducted in one day will be examined in this thesis. To examine slow perceptual learning, longer training regimes are needed to be undertaken. In practice, CI users may require longitudinal training process (6 months or more) to cope with their devices [126, 201].

3.5 Summary

Perceptual learning can be achieved via extensive exposure to perception tasks. It induces neuroplasticity, which can modify the neural map of the sensory cortex, and hence create a long-lasting effect that can be utilised to optimise perception of trained and novel stimuli. The application of perceptual learning has found its place in rehabilitative training which aims to optimise the listening experience of those receiving hearing aids or cochlear implants. Auditory training has proven to be effective in improving the perception of a variety of speech stimuli presented to hearing aids and CI recipients, and normal hearing listeners using CI simulation. The use of computerised auditory training has helped to implement the basic principle of auditory training and deliver improved auditory perceptual learning compared with classical training. A recent line of research promotes the use of audiovisual speech in auditory training, as evidence has been found for improved perceptual learning when visual signals were introduced in the auditory training. A link between the saliency of visual cues and improved outcomes of the audiovisual training has been found in the language training literature. Such links provide the motivation to investigate the impact of automatically increasing the saliency of visual speech presented in audiovisual training on improving audio-only listening skills. As a first step in this investigation, a roadmap to visual speech enhancement in this thesis needs to be drawn. Chapter 4 will introduce the visual speech enhancements as well as the

Audiovisual Training Framework that will be used to structure the evaluation of the impact of visual speech enhancement in audiovisual training.

Chapter 4

Visual Speech Enhancement: Methods and Evaluation Framework

4.1 Introduction

In a face-to-face speech chain model (Figure 2.1) in which a CI user is acting as a listener, the visual signals originating from the mouth (or visual speech) are highly weighted [295] and have a greater contribution to speech perception, since the auditory signal is degraded due to internal and/or external conditions (See Section 2.5.3). The question that arises in this context is: would increasing the saliency of the visual speech signals sourced from the corresponding production chain produce a consequent increase in their contribution to the CI perception chain? In other words, in an analogous way to auditory volume increase, is it possible to increase the intelligibility of the visual signal for CI users? A road-map to answer this question should consider the following:

1. For a given video in which audiovisual speech (i.e., the talker is being heard and seen) is presented, what are the enhancement methods that can be applied to the visual part of the speech, and to what extent can we enhance the visual signals, given that the corresponding auditory signals remains unenhanced?

2. As mentioned in Section 2.5.1, it is difficult to find a homogeneous CI user subject group for testing purposes, due to the variability in CI speech perception amongst users. Given that, is there an alternative speech perception chain model that can be used to predict CI users' reactions to visual speech enhancement?

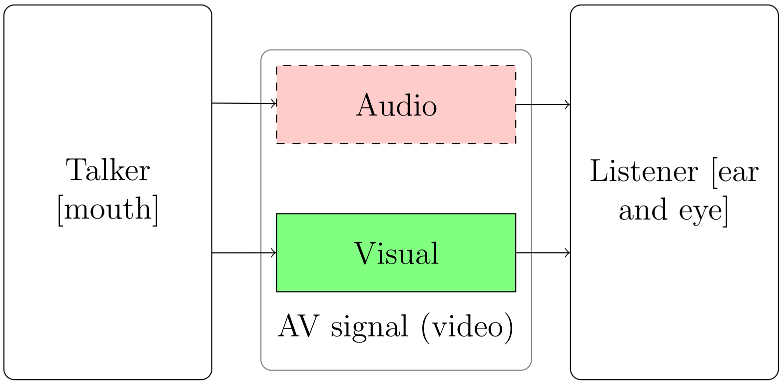


Figure 4.1: Perception speech chain of CI users. Green: A strong bias to the visual signal [295]; Red: distorted due to internal or/and external adversity.

3. What is the application/context in which visual speech information plays a special role and can be employed to test the enhancement effect?

In this chapter, a framework for visual speech enhancement is presented by addressing these points. Section 4.2 will present a review of visual speech characteristics. Based on this review, two visual speech enhancement methods are proposed: an appearance based method that improves the appearance of the external articulators, and a kinematics based method that increases the saliency of the mouth kinematics during speech production. Both methods will be demonstrated in detail in Chapters 5 and 6. In Section 4.3, non-native normal-hearing listeners will be considered as an alternative perception chain model to the CI users. This subject group will be treated as 'proxy' listeners that could predict the performance of the CI users when the proposed enhancement methods are applied. Section 4.4 will introduce the audiovisual training framework that is used in this thesis to evaluate the effectiveness of visual speech enhancement. This framework will be adapted in Chapters 5 and 6 to provide the intended evaluation for each method. Lastly, a pilot study that evaluates the effectiveness of the framework is presented in Section 4.4.1.

4.2 Visual Speech Enhancement

Although visual speech signals are not masked by external noise sources, they can be affected by a range of talker-dependent factors including lip emphasis [47, 173, 279], facial hair [166], speaking style [154, 279], and talker's gender [60]. For example, the size of the talker's lips can affect the clarity of lip movements; facial hair, such as a moustache or beard, can mask lip movement and alter facial configuration; an unfamiliar accent, and a lack of non-verbal cues and facial expressions can restrict the benefits of visual speech [154].

Evidence from the literature has also shown that increasing the visibility of the articulators can increase the benefit of visual speech. For example, Lander and Capek [173] compared the lip-reading performance of sixty normal-hearing listeners of a talker wearing bright red lipstick, no lipstick (natural), and concealer on the lips. Results showed that both applying lipstick and concealer, i.e., colouring the lips, improved the lip-reading performance of words and sentences compared to natural lips, indicating that increasing lip emphasis can improve the saliency of visual speech. Lips, tongue, and teeth visibility can also increase the benefit of visual speech [279]: Rosenblum *et al.* [266] and McGrath [210] found an increase in performance accuracy of 6% in listeners who were lip-reading a talking head when the teeth were added to the head.

The ability to regulate the effects of visual speech suggests the potential for investigating visual speech enhancement to aid audiovisual perception in adverse conditions. Given an audiovisual signal presented in a video format featuring a single talker with a frontal view of his/her full face, the notion of visual speech enhancement in this thesis refers to: an (off-line) process that aims to aid a listener's perception of receiving an audiovisual signal under adverse conditions by enhancing some articulatory aspects of the talker's speech production, so as to improve the visual signal's saliency. It is an off-line process, as the video of interest is post-processed in order to implement the enhancement technique. The proposed visual speech enhancement methods will be applied to the visual speech signal received by a perception chain in which the listener is experiencing adverse conditions (Figure 4.1). The enhanced visual signal is then coupled with the degraded audio signal (vocoded speech) to create visually-enhanced audiovisual signals to support the audiovisual perception of that listener.

It is vital to understand what constitutes visual speech in order to enhance it. According to Rosenblum [267], there are two classes of visual speech primitives that make visual speech cues: static cues and time-varying dynamic, or kinematic cues. Examples of static cues include the physical features of the articulators, such as lip spreading/rounding and tongue height [224] and the degree of lip opening [205]. These features can be considered pictorial, defined by 'demarcating of facial features' using the articulators' appearance characteristics [267]. For example, the articulators' texture colours [205, 224, 300] and luminance variations can help listeners to distinguish between mouth shapes. Shadow can also help to enhance the perception of the facial contours [50, 150], and shading provided by differences in illumination can inform depth cues [256]. The second class of visual information is kinematic cues such as lip and jaw motion trajectories [39]. Rosenblum *et al.* found that kinematic information that is derived from a point light technique can enhance the intelligibility of speech in noise [266].

Based on Rosenblum's classification, two enhancement methods will be explored in this thesis: *an appearance based enhancement method* that enhances the static cues of visual speech, and *a kinematics based enhancement method* that targets the dynamic aspects of visual speech. Using these methods, the extent of visual speech enhancement is also explored by investigating two extremes of visual speech enhancement: first, when the correlation between the auditory and the visual signal post enhancement is preserved, resulting in congruent audiovisual stimuli; second, when the congruency between audio and visual speech might be compromised due to the enhancement itself, resulting in mismatched audiovisual speech stimuli. The following sections will introduce each enhancement method.

4.2.1 Appearance Based Enhancement

In this method, the effect of automatically enhancing the appearance of the mouth on increasing the benefit of visual speech is tested. Since such an enhancement will target the appearance of the mouth, the correlation between the auditory signal and the visual speech will remain unaffected by the enhancement. Although the behavioural studies reviewed in Section 4.2 suggested a great potential for visual speech enhancement, computer-based methods are required to verify the effectiveness of these effects, especially for studies that involve testing talkers under different conditions in which psychological and environmental variables might interact and affect the speech production of those talkers. For example, although Lander and Capek [173] found that a lipstick effect increases the benefit of visual speech, they argued that talkers wearing lipstick may have produced some exaggerated gestures while talking, driven by the psychological impact resulting from wearing lipstick.

The proposed enhancement method uses the work of Lander and Capek [173] as a theoretical base. The method tests the effect of applying lipstick on the visibility of the mouth by automatically simulating a talker in audiovisual stimuli wearing lipstick. The effect of this enhancement is then evaluated against the unaltered-audiovisual and audio-only stimuli. Chapter 5 will cover the appearance-based enhancement (the simulated lipstick effect) in detail.

4.2.2 Kinematics Based Enhancement

Speech kinematics, or speech motion cues, is an aspect of speech that can be enhanced. Prospective enhancement techniques can either visualise or augment the kinematics cues. Ideas for visualisation are illustrated in Figure 4.2. For example, the magnitude of the facial change relative to the produced sound might be visualised using heat-maps superimposed on the talker's face; Richoz *et al.* [259] used a similar concept to visualise the emotional states of a talker on 3D heads for analysis purposes (Figure 4.2a). The articulatory movements can also be tracked using optical flow techniques (e.g. the Lucas-kanade optical flow method [195]) and then visualised to aid the identification of the produced sound (Figure 4.2b). Alternatively, the articulatory movements can be visualised based on knowledge of the produced sound by highlighting the muscle movements associated with that sound, irrespective of the movement made by the talker. This is to aid the perception of talkers with unintelligible speaking style (Figure 4.2c).

However, the introduction of artefacts on the talker's face might obscure some useful natural cues. It also requires training to learn to integrate these artefacts with the speech cue. Augmenting the kinematic cues by exaggeration is an alternative technique that may produce a more natural effect (Figure 4.3).

The exaggeration of kinematic cues is a type of motion signal processing, which is an established field that combines signal and image processing techniques to implement motion curves manipulation in a given scene [40, 188, 325, 332]. For example, Unuma *et al.* [318] interpolated and extrapolated motion data to animate a 3D character; Wang *et al.* [325] convolved the motion signal with an inverted Laplacian of a Gaussian filter to exaggerate the motion in a given scene (e.g. a walking character) and then applied a deformation technique on the video to show the exaggeration effect; Theobald *et al.* [309] addressed speaking-style exaggeration in 2D videos in order to support forensic lip-reading of surveillance videos; they exaggerated the lip movements of a talker in a video by amplifying the talker's mouth shapes and appearance. They found improved lip-reading performance among inexperienced lip-readers (the effect was not tested when combined with the audio signal). Without any manipulation to the speech signal motion, Alexanderson and Beskow [13] used an intrinsically exaggerated speech signal (visual Lombard speech) to animate a 3D avatar. They found increased intelligibility in visual Lombard animations combined with auditory plain speech, and when combined with Lombard speech. This is compared with animations driven from visual plain speech data and combined with auditory plain speech (See Section 2.5.4).

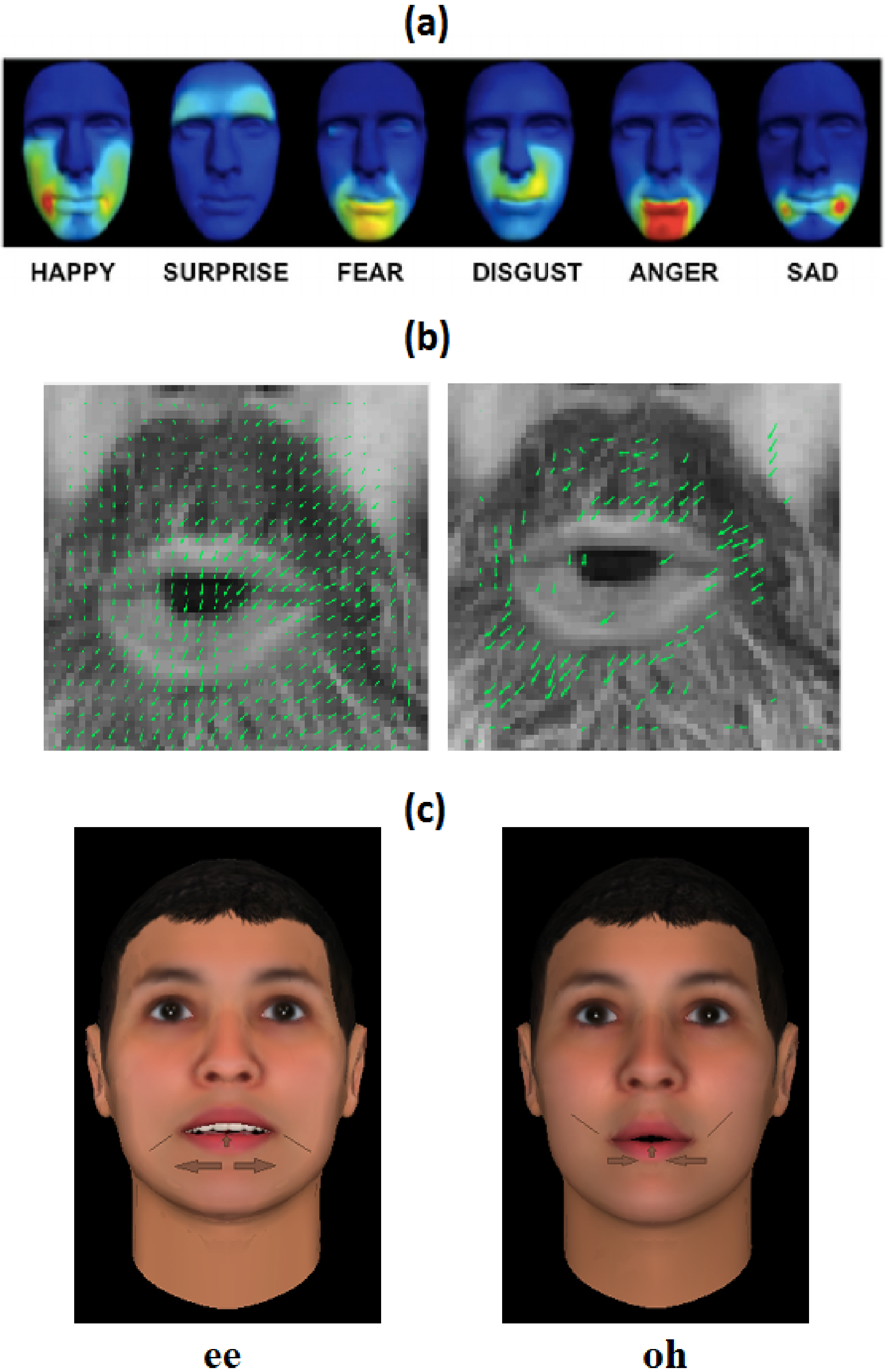


Figure 4.2: Alternative visualisation approaches of speech kinematics: (a) A heat-map is superimposed on the talker's face; inspired from visualising emotion using heat-maps, image by Richoz *et al.* [259] (permission to use the figure is granted); (b) Using optical flow to track changes in face movements (c) Labeling the movement of the mouth that is associated with the produced sound. The 3D head model is generated by using FaceGen [145].

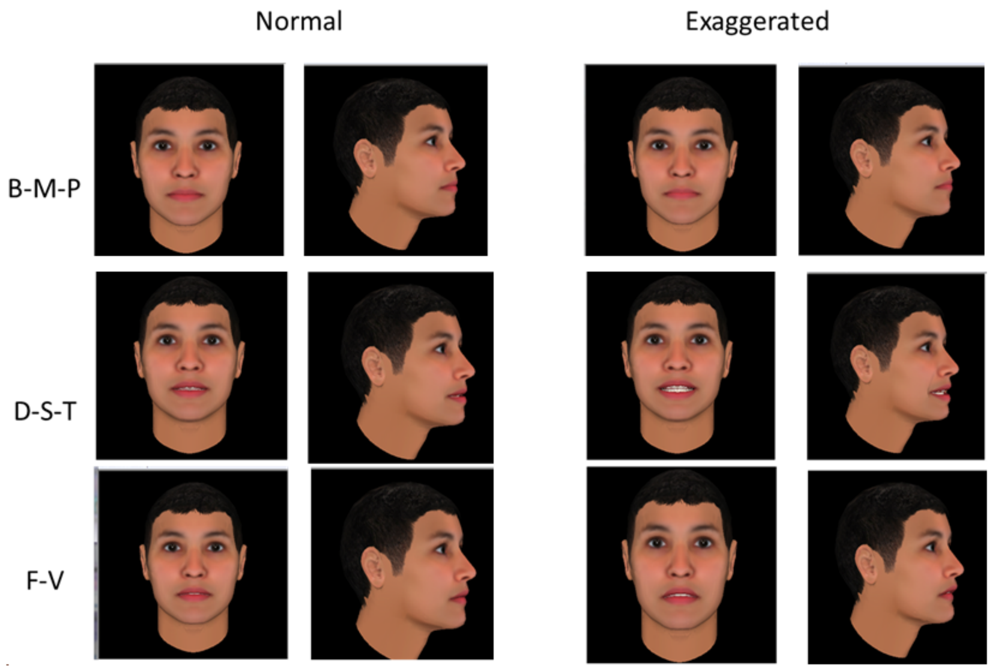


Figure 4.3: Speech kinematics augmentation: a prototype for exaggerating facial movement. The 3D head models generated by using FaceGen [145].

Based on the ideas from the previous work, this thesis will investigate the exaggeration of kinematic visual speech cues. Its impact on the visual benefits of audiovisual speech will be tested. The main challenge in such an enhancement is that exaggerating only the visual signal in audiovisual speech could create conflicting, incongruent audiovisual inputs for listeners. The impact of exposure to conflicting inputs has been widely investigated in the behavioural-studies literature [26, 30, 67, 211, 273]. These impacts can be classified as immediate biases or after-effects [30]. An example of immediate biases may be observed in spatial conflict (such as the ventriloquism effect), where visual stimuli can influence sound localisation [26], and in identity conflict (i.e., the McGurk effect) [211]. Another study also found that exposure to mismatched inputs can create an after-effect on perceptual modalities used for adapting to this conflict. For example, visual speech has the ability to recalibrate auditory perception after exposure to the conflicting audiovisual stimuli observed in the McGurk effect [30]. Audiovisual recalibration has also been observed in temporal audiovisual conflicts (such as in live broadcasts) to help adapt to time lags [99, 323].

Given these factors, the kinematic-based enhancement method has two aims: first, to investigate the after-effects of listeners' exposure to the conflict between the articulation energy in the visual signal and the vocal effort in the acoustic signal (as well as determining whether the listeners will acquire the ability to adapt in order to overcome the audiovisual mismatch); and second, to study the impact of such modifications to audiovisual speech on improvements to audiovisual speech perception.

The exaggeration method that is of interest in this thesis is described as follows: for a given audiovisual speech video, the speaking style of the talker is exaggerated using limited knowledge of his/her speaking style (i.e., motion data of the talker from that video), and without any prior knowledge of the exaggeration style of that talker (i.e., motion data of real exaggerated speech produced by that talker). To achieve these goals, a similar exaggeration method to Theobald *et al.* [309] will be used to model mouth shape variation. Theobald *et al.* used the principal components analysis to model the shape and the appearance of the talker's mouth shapes. In the proposed method in this thesis, the mouth motion is exaggerated by extrapolating the PCs of a talker's mouth shapes in a given video. 2D image warping is then applied to reanimate the video using the new exaggerated mouth motion. 2D image warping, which involves the geometric transformations that define a relationship between two images' pixels [51], is a well-known technique for facial modifications that is used for visualising plastic-surgery outcomes and in various entertainment platforms [167, 183, 214]. Chapter 6 will present the kinematic based enhancement: the exaggeration effect. A combined enhancement of the lipstick effect (Section 4.2.1) and the exaggeration effect will also be examined.

This thesis also examines a real-life case study of visual speech enhancement observed in a hyper-articulated speech (the Lombard speech). Lombard speech features a collection of acoustic, phonetic, and articulatory modifications resulting from speech production modifications in order to counter adverse conditions. Increased intelligibility in Lombard speech is not only sourced from the acoustic and phonetic adaptations, but also from the articulatory adaptations [13, 63, 92, 161] (a detailed review of Lombard speech is presented in Section 2.5.4). Visual Lombard speech is addressed in this thesis as an example of visual hyper-articulation; it was chosen since it can be induced in a controllable manner [284].

The global change in visual Lombard speech adaptations has received much attention from researchers (see Section 2.5.4)); however, little attention has been paid to changes at the phoneme level [100, 102]. Moreover, such studies have addressed phoneme production adaptations in very limited contexts. In this thesis, the visual Lombard speech adaptations at the utterance-level and phoneme-level, and under different contexts, will be examined and characterised. As the literature lacks

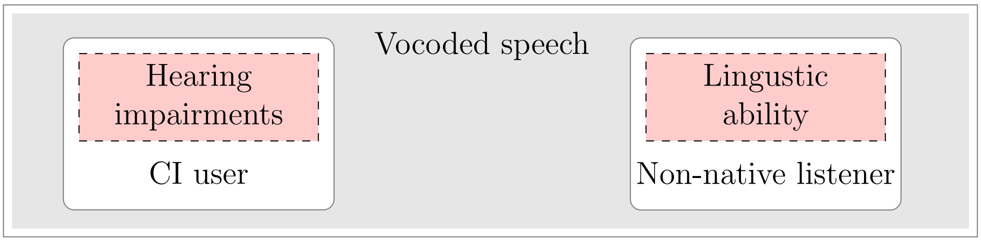


Figure 4.4: Perception models for CI users and non-native listeners. Red rectangle: internal adversity; gray rectangle: external adversity.

Lombard speech datasets suitable to carry out the intended analysis, a novel dataset of audiovisual Lombard recordings will be collected. The dataset, which is based on widely used audiovisual Grid corpus [54], will be recorded under high SNR level, whereas listeners are exposed to background noise through headphones presented at 80 dB SPL, and will offer a plain reference to each recorded Lombard sentence. It will also feature two synchronised views of the talker: a front view and a side view, which offers the chance to characterise visual Lombard speech from different angles. The video recording will be made using head-mounted cameras to stabilise the talker's head throughout recording, therefore allow precise comparison of the Lombard and plain utterances. The collection of the audiovisual Lombard dataset and the subsequent analysis will be presented in Chapter 7.

4.3 Alternative Perception Chain Model: Non-native Subjects as Listeners

The target users for the visual speech enhancements are hearing impaired people, in particular, CI users. As mentioned in Section 2.5.1, the main obstacle that faces CI researchers is the limited access to a homogeneous subject group for testing. This is due to the wide variation in implant outcomes amongst users, which hinders the full utilisation of the CI benefit. Such variation is also regarded as internal adversity in CI users. To form a more controlled group, many CI researchers in the literature have used normal hearing listeners in conjunction with CI simulation, i.e., test normal hearing listeners' reactions to speech processed using CI processing techniques. Yet, this doesn't reflect the hearing experience of CI users who also cope with internal adversity.

A possible way to model the effect of internal adversity in CI users is to look for an analogous group of normal hearing listeners whose perception chain model is affected by a source of internal adversity. One such group is non-native, normal hearing

		Bilabial	Labiodental	Dental	Alveolar	Post alveolar	Velar	Uvular	Pharyngeal	Glottal
Plosive	E	p b			t d		k g			ʔ
	A	b			t d tˤ dˤ		k	q		ʔ
Fricative	E		f v	θ ð	s z	ʃ ʒ				h
	A		f	θ ð	s z sˤ zˤ	ʃ		χ ɣ	ħʕ	h
Trill	E									
	A				r					
Approximant	E				ɹ					
	A									

Table 4.1: English vs. Arabic IPA chart. (Adapted from Binturki [31]).

listeners, who a number of researchers have found a source of internal adversity that affects this group [93, 294, 319]. As mentioned in Section 2.5, the perception of both CI users and non-native listeners is a function of internal adversity [93, 181, 208, 237, 246] external adversity [106, 175, 290], and visual speech [74, 80, 131, 132]: increased levels of internal and external adversity deteriorate the intelligibility of the perceived speech, and the introduction of visual speech can aid perception. Figure 4.4 illustrates the perception chain models for CI users and non-native listeners that are of interest in this thesis: the existence of the internal adversity in CI users (hearing impairment factors [237, 246]) and non-native listeners (linguistic ability factors [209, 212, 319]), and the external adversity represented in the vocoded speech (CI simulated speech). Non-native subjects can therefore be considered as analogous to CI users, since both types of listeners cope with internal and external adversity in their perception of CI-processed speech.

Unlike CI users, the provision of an adequate number of non-native listeners for testing is feasible as well as offering control of the subject group's homogeneity. In this thesis, the non-native listeners are non-UK-native, female Saudi-Arabian nationals attending the Department of Information Technology at King Saud University in Riyadh, Saudi Arabia. Table 4.1 shows the overlap between Arabic and English phonemes indicating a shared number of phonemes between the two language [31]. The Arabic language, however, lacks /p/ and /v/ phonemes, which, in consequence, results in a language transfer effect on audiovisual perception of the phonemic contrast under those categories [144].

The effect of the linguistic knowledge is controlled by considering a number of measures; one is by using the IELTS [International English Language Testing System] test score (in particular the listening band aspect) as a measure to select subjects. Another measure is taken in Chapter 6, in which a pre-listening test is used to divide listeners into balanced subgroups. Moreover, subjects share a similar educational background (all attend the Information Technology Department where textbooks, examination, and teaching materials are in English) and received the same level of English education in schools (all attended a government school in which English teaching starts in year seven, and there is a foundation year in English at university). Controlling the effect of linguistic ability is not only important for auditory perception, but also to control the sensitivity and the weighing of visual cues by non-native listeners [125]. To control gender differences in audiovisual perception, female subjects were chosen; females have been found to be more sensitive to visual speech than males and are better speech-readers [61].

After selecting the enhancement methods and the subject group, the next step to address is the application in which the visual speech enhancement is applied. An example of an application that can facilitate access to both auditory and audiovisual perception of a CI user is audiovisual training. Audiovisual training can shape the listening experience of CI users [97], and the provision of visual speech during the training can improve auditory and audiovisual perception [28, 130]. Moreover, a link between visual speech saliency and improved training outcomes was reported by Hazan *et al.* [131, 132] in which they found an increased learning gain by language learners in visually-salient phoneme categories among other phoneme categories (see Chapter 3). This evidence suggests that there is an impact from audiovisual training in improving perception. In this thesis, audiovisual training is used in assessing the impact of visual speech enhancement on improving the benefit of visual speech in the training stimuli, and the impact on auditory-only perception after the training. The following section will introduce *the Audiovisual Training Framework*.

4.4 The Audiovisual Training Framework

The theoretical framework that guides the audiovisual training introduced in this thesis is inspired by the work of Bernstein *et al.* [28], who found a link between introducing visual speech in auditory training (AV training) that aims to improve auditory skills and inducing CAS plasticity. Bernstein *et al.* [28] used vocoded speech to train two groups of normal-hearing native listeners, where each was assigned to a

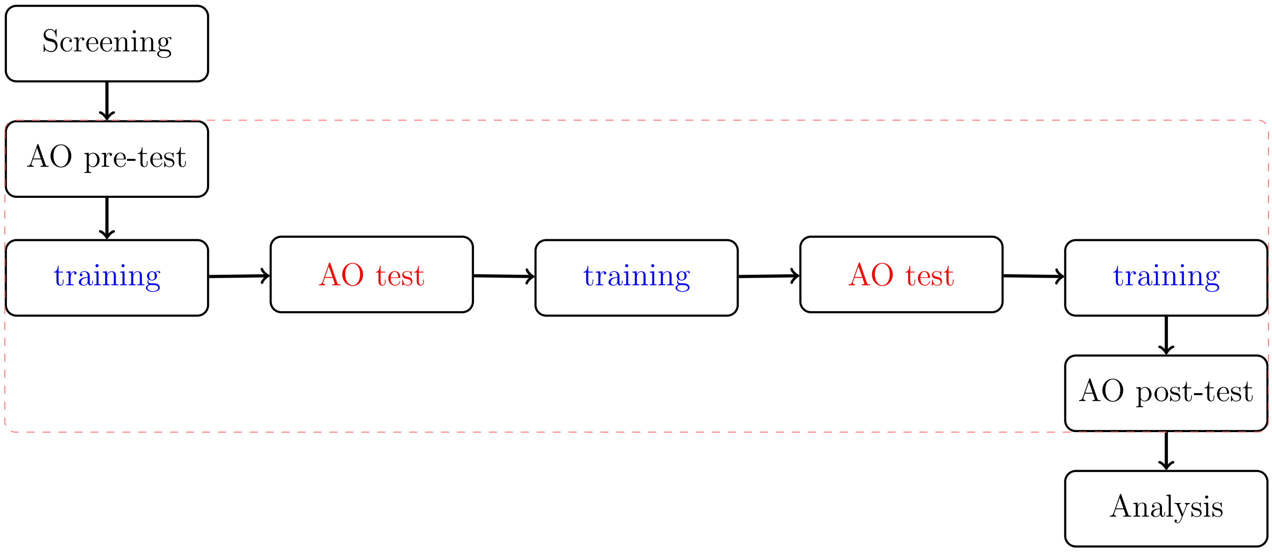


Figure 4.5: The audiovisual training framework. AO: auditory only.

different training modality (auditory or audiovisual), and then underwent three blocks of training bounded by audio-only pre- and post-tests. These tests aimed to quantify the training gain on improving audio-only skills, or auditory perception. This thesis adapts Bernstein' s framework with some modifications, listed as follows:

- *Training duration*: Bernstein *et al.*'s framework repeated the training over four days with the aim of testing slow perceptual learning. In this thesis, the training process was shortened to be a one-day experiment addressing the rapid impact of perceptual learning on improving auditory perception (See Chapter 2 for more information about types of perceptual learning impact).

- *Subjects group*: Bernstein *et al.* used native listeners, while this thesis uses non-native listeners, as described in Section 4.3, to evaluate the effect of the training.

- *Stimuli type*: in Bernstein *et al.*'s framework, 'nonsense' training stimuli were employed to control the impact of lexical information on guiding perceptual learning [28, 65]. However, in order to activate the effect of using non-native listeners in the experiments in this thesis, the nonsense stimuli are replaced with meaningful sentences that convey lexical information.

- *Training milestones*: short audio-only tests following each training sessions are introduced. These assess the speed of learning and mark training milestones for each training group.

command	colour	preposition	letter	digit	adverb
bin lay place set	blue green red white	at by in with	A-Z excluding W	1-9, zero	again now please soon

Table 4.2: Sentence syntax for the Grid corpus. (Adapted from *Cooke et al.* [54].)

This framework evaluates the impact of using visually-enhanced audiovisual speech in the audiovisual training of CI simulated speech. Figure 4.5 summarises the main processes in the training framework. According to Erber's classification of the auditory training process [86], this audiovisual training framework is in the discrimination and identification categories and follows the synthetic training approach (see Section 3.3).

The baseline training modalities that will be used to guide the evaluation comparison are the audio-only and the un-enhanced audiovisual training modalities. The training framework can therefore be adapted to provide auditory training under n training modalities to n subgroups where $n > 2$. Before using the audiovisual training framework in the following chapters, it is important to evaluate the effectiveness of this framework. In the following section, a pilot study that evaluates the impact of using a visual signal on the audiovisual training results using the proposed audiovisual training framework is presented. The evaluation will address two main points:

- The effectiveness of the proposed audiovisual training framework. As mentioned, the framework replicates, with adaptations, Bernstein *et al.*'s framework [28]. Thus, the effect of the adaptations on the audiovisual training gain needs to be examined.

- An evaluation of the alternative perception chain model (i.e., the non-UK native, Saudi listeners). This is done by comparing responses from normal hearing native and non-native listeners to the audiovisual training framework.

4.4.1 Evaluation

Training Stimuli

Training sentences were extracted from the Grid corpus [54]. The Grid is an audiovisual database designed for computational-behavioural research, and originally inspired by the auditory-only coordinate response measure (CRM) corpus [228]. The format of a CRM command is READY [call sign] GO TO [colour][digit] NOW. The

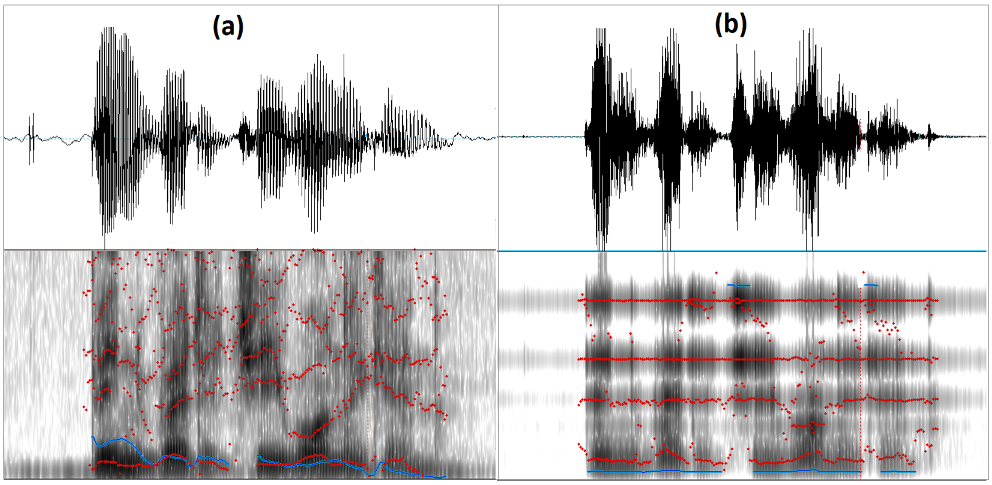


Figure 4.6: The spectrogram of (a) original and (b) vocoded Grid sentence: Bin red at G 1 again. Red lines: formants, blue lines: pitch contour (Generated by Praat [32]).

CRM corpus was populated by eight call signs, four colours, and eight digits, which resulted in 2048 stimuli of multiple talkers. The Grid corpus modifies the format of CRM's stimuli with richer high-level semantic details, by including four commands, four prepositions, 25 letters, ten digits and four adverbs (Table 4.2). Accordingly, the possible permutations generated from these keywords are 64,000 sentences; the corpus includes audio and video recordings of 34,000 sentences from these permutations, collected from 18 male and 16 female talkers (34 total, i.e., each talker uttered 1000 sentences). Experiments on Grid corpus sentences have shown high intelligibility in quiet and noisy environments [54]. There are many audiovisual speech corpora that can be used to provide training stimuli with high level lexical information such as CUAVE [243], XM2VTSDB [219], and VidTIMIT [275]. The Grid, however, was considered a convenient source for training stimuli since the structure of the sentences features the following:

- Consistent syntax in all sentences;
- The provision of keywords (i.e., the colour, the letter, and the digit) that can be used as the identification task;
- Rich phonetic features due to the inclusion of the alphabet;
- Simple sentences in a short format, which helps to diminish the effect of using memory, and a lexicon of familiar words.

24 AO pre-test → **Session1:** 20 X, 6 AO → **Session2:** 20 X, 6 AO → **Session3:** 20 X → 24 AO post-test

Figure 4.7: The training method consists of an audio-only pre-test followed by three training sessions then an audio-only post-test session for determining the training gains.

A pool of 250 Grid sentences of a selected talker (ID= 1) were randomly chosen to ensure that listeners provided coverage of the same sentences. This pool was then randomly split into ten sets of twelve sentences, ten sets of ten sentences, and five sets of six sentences to be used in the training framework.

Two sets of training stimuli were created to be used in the audiovisual training framework evaluation: audio-only (AO) and audiovisual (AV). To create the AO stimuli, the audio tracks of the selected sentences were altered using an eight-channel sine wave vocoder (AngelSim [96]). The sine wave vocoder was used instead of a noise vocoder as it simulates the output from a CI processor without introducing noise [24, 289]. Eight channels were used in the simulation, as normal hearing listeners when listening to an eight-channel simulation and users with an eight-channel CI were found to achieve a comparable level of speech understanding [79]. The vocoding process started by applying a bandpass filter that divided the input signal into eight-channel between 200 to 7,000 Hz (slope = 24 dB). Each channel was then low-pass filtered by 160 Hz (slope = 24 dB) to obtain the envelope. The envelope of each channel then modulated a sine wave that replaced the signal frequency. Figure 4.6 shows the spectrogram of the vocoded and original grid sentences. The vocoded spectrogram features vertically compressed and horizontally expanded formants and a flattened pitch contour compared with the normal recording. Such modifications are likely to affect the intelligibility of vowels, as well as the accessibility to voicing and intonation cues [289]. To create AV stimuli, the Grid videos (showing the talker's face) of the selected sentences were processed using FFMPEG [22] to replace the audio Grid sentence in each video with the vocoded ones (AO stimuli).

Procedure

Prior to conducting the training, the hearing of a subject was screened using an on-line pure tone audiometric test [249]. The tones used in the test were calibrated against an audiometer and are based on the international standard ISO3897:2005 (which specifies a reference threshold of hearing for the calibration of audiometric equipment

[37, 38, 235]), recommended by the British Society of Audiology. The screening threshold was 25 dB HL for frequencies between 125 Hz to 8 kHz. The subject viewed their screening results using an audiometer printout provided by the test. The researcher explained to the subject how to read the results from the audiogram. The subject continued the experiment but their data was excluded if she failed to attain the minimum threshold for all frequencies. In this case the subject was also advised to have a hearing test conducted by a professional audiologist.

Figure 4.7 illustrates the training procedure. In order to set a baseline level for each subgroup, all subjects took an audio-only pre-test of 24 AO stimuli (12 stimuli repeated twice), presented in random order. Subjects then underwent three training sessions. Each session consisted of a training block that used ten X stimuli (X = AO or AV), repeated twice, then presented in a random order to give 20 X, and a test session of six AO stimuli (except for session 3, which preceded the post-test). The two sets of six AO presented in sessions 1 and 2 were used to track learning milestones for all subgroups. After completing all training sessions, all subjects took an audio-only post-test using 24 AO stimuli (12 stimuli repeated twice and presented in a random order) to assess their training gain.

A training tool, developed in C#, was used to present training and testing stimuli and collect responses from the subjects. A profile for each subject was created first by the training tool, which contained demographic information such as age band, gender, the subject's first language and English proficiency level. The tool then generated a unique identification number for the subject, and assigned him/her randomly to a training modality. Also, the tool assigned to each subject seven random sets of sentences: two sets of twelve sentences, three sets of ten sentences, and two sets of six sentences in order to be used in the testing and training sessions. The sets of twelve were used in the pre- and post-tests, while the three sets of ten and the two sets of six were used in the training process.

The interface of the training consists of a stimuli viewer, a track bar for AO stimuli or a media viewer for AV stimuli. The subject task was to identify three keywords: colour, letter and digit, that conform to the played stimulus. The subject used a keyboard to enter the keywords in three text-boxes specified under the stimulus viewer. Four colour-coded keys on the keyboard were allocated for colours, as well as the number keypad for digits, and the alphabet keys for letters. During training (i.e., 20 X in sessions 1, 2 and 3), after the subjects submitted their input for a given stimulus, that stimulus was then replayed with added subtitles to show the correct words, whether or not the input was correct. No such feedback was provided during testing (i.e., during the 24 A pre-test, the 6 A in session 1 and 2, and the 24 A post-test). A response was considered correct if the subject successfully entered all the required keywords; however, the system kept track of the successful responses for each keyword. All subjects' responses were recorded and saved by the tool. The training tool hence provides the two main principles for perceptual learning: the repetitive exposure to stimuli, and the provision of feedback (See Chapter 3).

Subjects

Two groups of normal hearing participants were recruited in two different geographic locations. The first location was the female campus of King Saud University in Riyadh, Saudi Arabia (12 non-native Saudi listeners, IELTS score $\in$ [5.5,6]). The second location was the Department of Computer Science, University of Sheffield (9 native English listeners). All subjects were in the age range 18–30 (M = 24 years, SD = 4.5 years). The Saudi participants were subgrouped into equal groups, A_s and V_s, and the English participants were sub-grouped into groups of 4 and 5 participants, A_e and V_e. Where A and V denote audio-only training and audio-visual training, respectively, and the subscripts s and e denote Saudi and English listeners, respectively. Ethics permission for this study was obtained by following the University of Sheffield Ethics Procedure.

Results

Figure 4.8a shows the sentence recognition scores during the training for all subgroups. These scores were used to provide a subjective intelligibility assessment [292] of the speech used in the training. A clear gap is observed between the Saudi and the English subgroups in recognition scores across sessions, except for session 2 in which the A_e and the V_s subgroups showed comparable results. There was no significant difference between the A_e and the V_e subgroups; this is perhaps due to the impact of the lexical cues presented in the training stimuli that might have served the same supplementary role as the visual cues [28]. The lexical cues support, however, was not evident within the Saudi group; the V_s subgroup outperformed the A_s subgroup in the recognition scores, possibly due to the role of the visual cues for this population.

Figure 4.8b compares the pre- and post-training scores for the auditory only tests. The V (V_s and V_e) subgroups achieved a higher training gain than the A (A_s and A_e) subgroups. Moreover, the V_s subgroup's performance in the post-test reached a comparable level to the baseline level of the English subgroups (A_e and V_e). Whilst the A_s subjects improved, they were not close to the baseline levels of the English

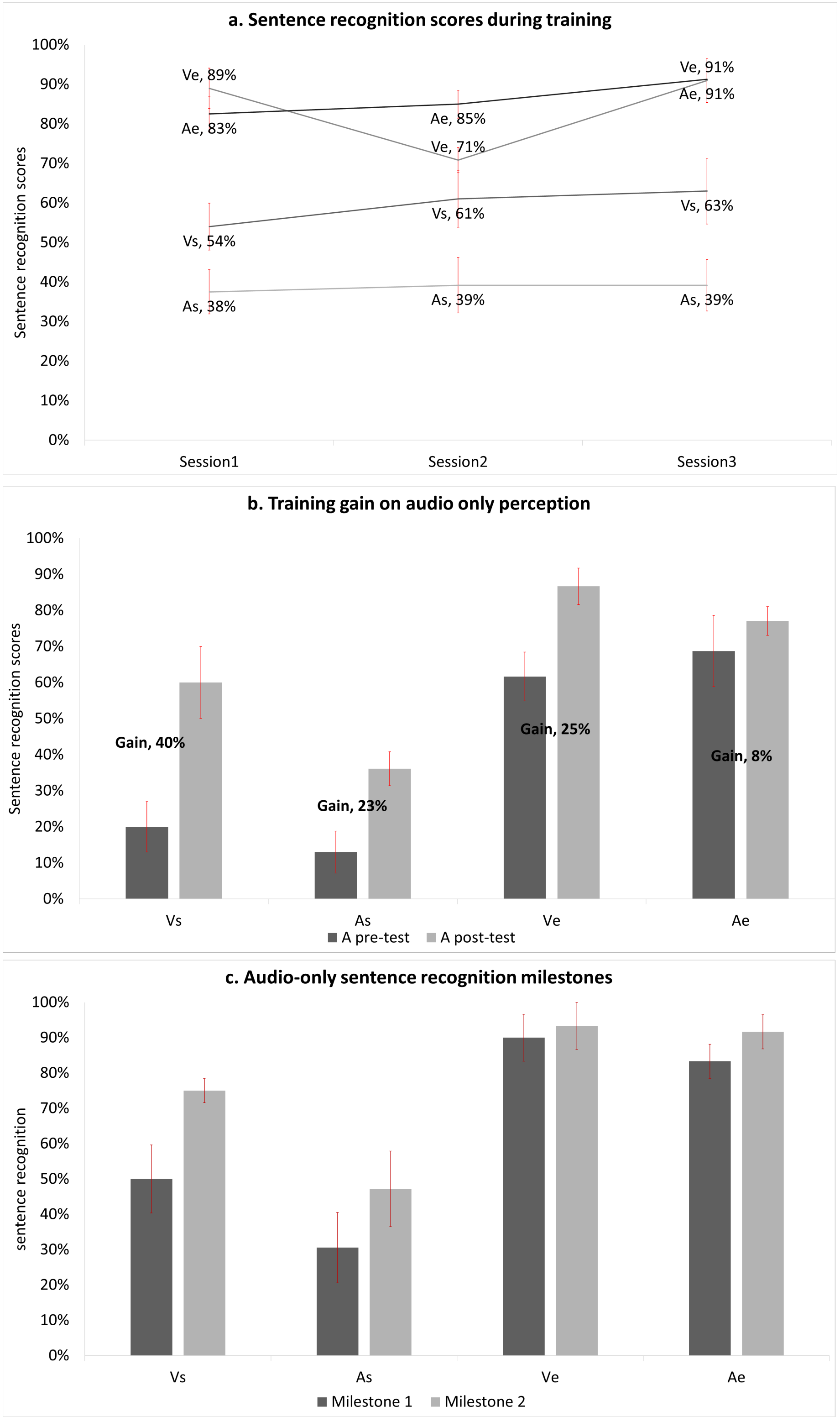


Figure 4.8: Results for the A_s, V_s, A_e, and V_e subjects: (a) Sentence recognition during training; (b) Audio-only pre- and post-test mean identification scores and training gains (post-test results and pre-test results); (c) Training impact on audio-only sentence recognition (learning milestones). Errors bars =/- standard error.

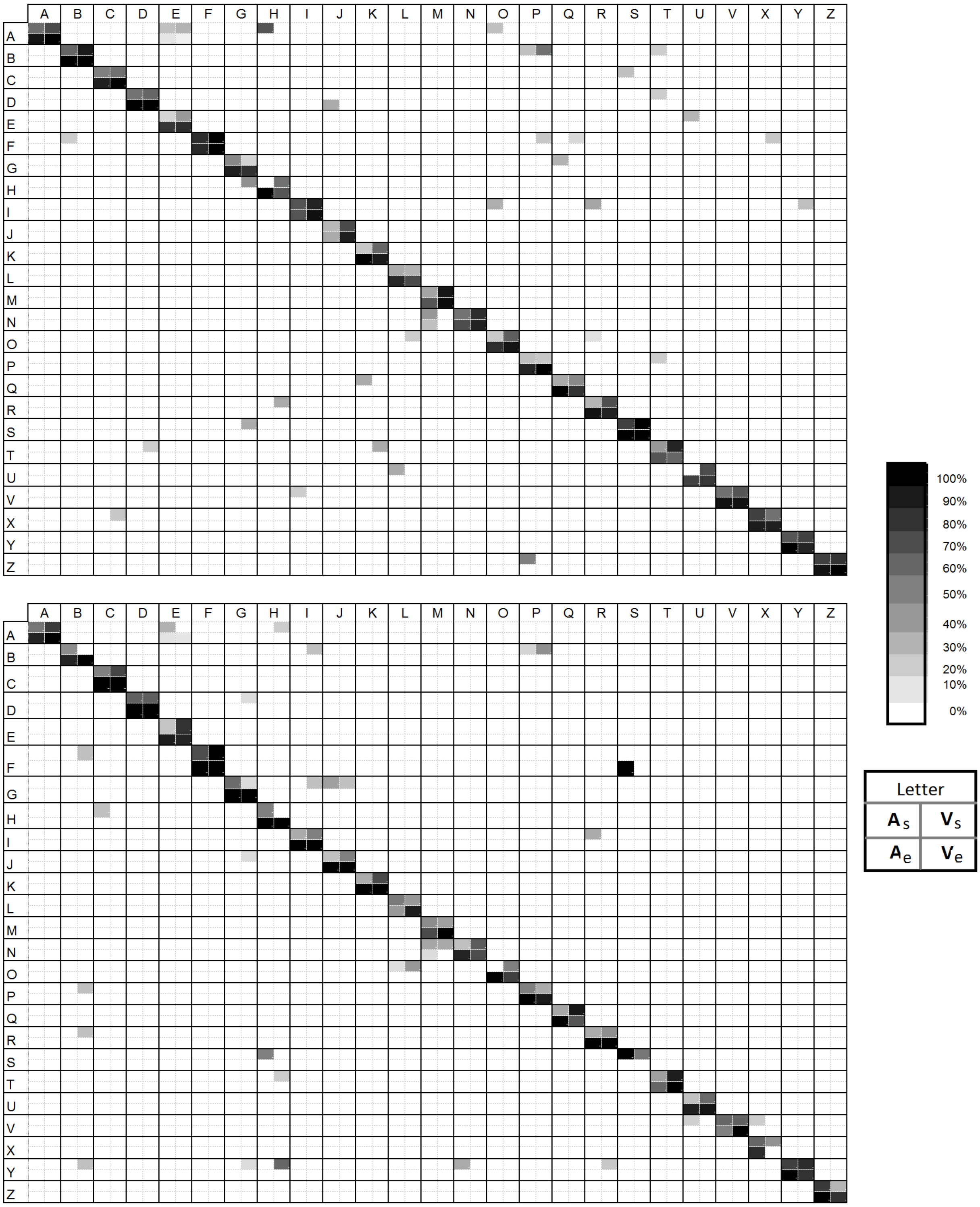


Figure 4.9: Confusion matrix of letter-recognition scores during training (top) and audio-only post-testing (bottom) for A_s, V_s, A_e, and V_e. Column: actual letter; row: listeners' response; Diagonal: letter recognition mean rates; Elsewhere: confusion mean rate for respective row-column pairs. Colour shades: the scale of recognition/confusion mean rate, the darker the shade, the higher the response to this cell.)

subgroups. One point to mention is that the V_s subgroup performed better than the A_s subgroup in the pretest, suggesting that they were a better subgroup of listeners. This makes the post-test harder to interpret since the V_s subgroup might have shown greater improvement because they started with better performance in the first place.

Figure 4.8c shows the scores for the stimuli used in the 6 AO test of sessions 1 and 2 for all subgroups. These scores were used as learning milestones to track the audio-only skills for all subgroups throughout the training. Unlike the other subgroups, the V_s subgroup showed significant difference between milestones 1 and 2. This could be an indication of the impact of using a visual signal in accelerating the perceptual learning of auditory skills.

Confusion matrices (Figure 4.9) were also produced in order to understand the possible sources of confusion the subjects experienced in letter keywords recognition during the training and audio-only post-testing. Letter recognition was found to be the most challenging task for all subjects due to the need to select from a larger set with high variance (25 letters). Letters also have shorter duration in terms of phonemes, and thus less information, as opposed to colours (4) and digits (10). The diagonal represents the recognition mean rates for letters. Scores elsewhere represent the confusion mean rate for respective row-column pairs. Colour shades represent the scale of recognition/confusion mean rate (the darker the shade, the higher the response to this cell). The strong main diagonal pattern in Figure 4.9, observed for the V_e and the A_e subgroups, clearly illustrates that the English subgroups were less confused than the Saudi subgroups.

For the English subgroups, no significant difference was spotted between V_e and A_e. For the Saudi subgroups, V_s showed better overall performance in letter recognition, especially in the post-test results. This was confirmed by t-test result that showed a significant difference in letter recognition during post-test (T = 2.63, P = 0.013) between V_s and A_s. The V_s subgroup achieved higher scores in the identification of letters that are constructed from diphthongs (a, e, i, u) with a significant difference ($T = 3.12$, $P = 0.02$). Since vowels differ in the frequency of the first formants (F1 and F2), and given that F1 and F2 are correlated with jaw height and tongue position [190], visual signals may contribute to enhance the intelligibility of diphthongs by the V_s subgroup, while the A_s subgroup showed high confusion between A and E (confusion mean= 31%) and I and O (confusion mean= 13%). The V_s subgroup also outperformed the A_s subgroup with a significant difference (P = 0.002) in identifying the nasal sound in N and voiceless plosive sound. On the other hand, the V_s subgroup showed confusion between visually similar letters, for example

G and D, and P and B, compared with the A_s subgroup. In these letters, visual signals may have impeded learning the invisible sounds (such as postalveolar and velar) that were bounded by visible ones (such as vowels and alveolar), for example, the invisible sound /ʒ/ in G /dʒiː/.

As the voicing information is generally affected by the vocoding process, all subjects reported difficulty in discriminating between voiced B and voiceless P. The Saudi subjects were more affected due to a language specific factor, given that the Arabic phonemic inventory lacks the voiceless sound P. However, in the V_s subgroup, the confusion rate was higher (confusion mean = 44%), indicating that visual cues may have impeded the learning of voicing discrimination of the pair B and P.

Discussion

This pilot study has evaluated the audiovisual training framework's impact on enhancing auditory only speech perception. It has also examined the nativeness effect as a source of internal adversity by comparing the training effect on two groups of normal hearing listeners: native and non-native. The main observation and discussion points are as follows.

The Audiovisual Training Framework is Effective The audiovisual training framework was effective for all subgroups, but, in different magnitudes. V subgroups attained higher training gain than A subgroups, consistent with previous studies that confirm the effectiveness of audiovisual training [28, 250]. The effectiveness of audiovisual training was also reflected in the fast learning pace observed for the V_s subgroup.

Speech Intelligibility During the Training A prominent difference between the native and non-native listeners' responses to the vocoded speech during the training is noted, an observation that aligns with the reported literature on the effect of nativeness in speech perception under adverse conditions [106, 175, 209, 319]. The benefit of introducing the visual speech was significantly higher in the Saudi listeners than in the English listeners; using Sumby and Pollacks' [296] metric, the visual contribution to speech recognition under noisy conditions can be quantified as follows:

$$C_V = \frac{C_{AV} - C_{AO}}{1 - C_{AO}} \tag{4.1}$$

where C_{AO} and C_{AV} is the normalised recognition scores of AO and AV, respectively. Therefore, the benefit of introducing the visual signal on the Saudi and

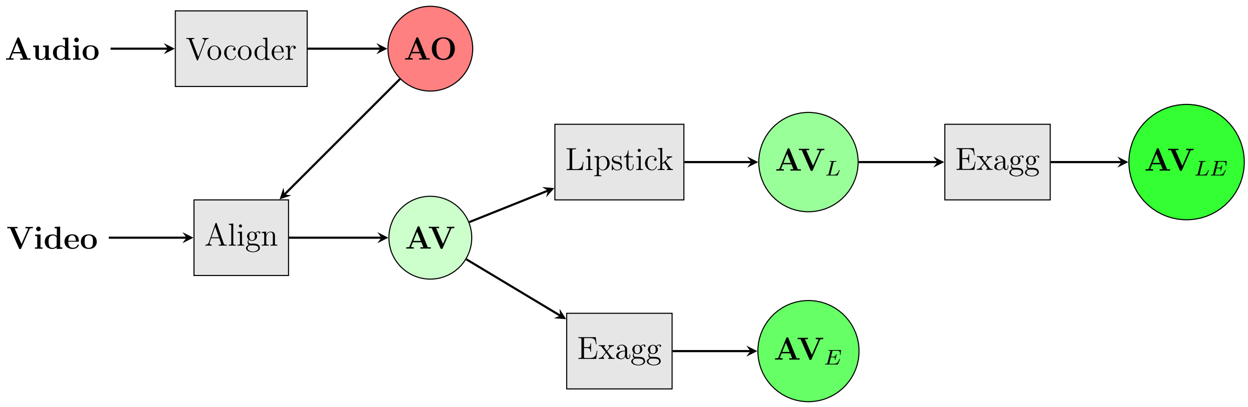
English listeners' speech perception (using scores from the last training session) is 0.39 and 0.00, respectively. Visual signals also seemed to slightly improve the transfer of some nasality and voicing information as reflected by the letter confusion matrices. These observations are consistent with previous finding on the benefit of visual speech to non-native listeners [130–132]. These observation, the increased deterioration in perception under external adverse conditions and the increased benefit of visual speech, are also observed in CI users [74, 226] (see Section 2.5.1). This may support the hypothesis that using non-native subjects as listeners might help to consider the difference between the native normal hearing individuals and CI users' responses to CI processed speech.

V_s Achieved Higher Gain Than $\mathbf{V}_e$ In the case of native listeners, the lexical cues provided within the training stimuli are possibly responsible for guiding the perceptual learning [28]. Non-native listeners, however, may have had limited access to the lexical cues due to internal adversity, and hence might utilise visual speech cues to guide more effective perceptual learning. It is unclear, however, whether V_s adaptation was at the level of the internal or the external adversity. One observation suggests that it could be at the internal level: after the training, the V_s subgroup reached a comparable level to the baseline state of the English subgroups who are internal-adversity free.

Given these results, the audiovisual training framework will be used in the evaluation of the visual speech enhancement methods in Chapters 5 and 6, with some modifications. Table 4.3 illustrates the stimuli and training modalities that will be used by the audiovisual training framework in Chapters 5 and 6. In Chapter 5, a new audiovisual stimuli, AV_L, will be created by applying the lipstick effect to AV stimuli, and the training framework will be adapted to introduce three training modalities: A, V and E_1. In Chapter 6, the audiovisual training framework will be adapted to introduce two new stimuli: AV_E and AV_{LE} that are created by applying the exaggeration effect to AV and AV_L stimuli, respectively. The audiovisual training framework will then be modified to introduce four training modalities: A, V, E_2 and E_3.

4.5 Summary

This chapter presented a framework for visual speech enhancement in this thesis. Two methods of enhancement were introduced: appearance based and kinematics



Stimulus	Training	
AO	A	Auditory training
AV	V	Audiovisual training
AV$_L$	E$_1$	Enhanced audiovisual training - lipstick effect
AV$_E$	E$_2$	Enhanced audiovisual training - exaggeration effect
AV$_{LE}$	E$_3$	Enhanced audiovisual training - lipstick and exaggeration effect

Table 4.3: Training stimuli and modalities.

based methods. The lipstick effect will be implemented as an appearance-based enhancement method. According to Lander and Capek's [173] observations of talkers who wore real lipstick, this effect can improve lip-reading. For kinematics based enhancement, the effect of visual speech exaggeration will be implemented; this effect was found to be effective by Theobald *et al.* [309] on inexperienced lip-readers. The two methods will also examine the extent to which visual speech enhancement can be applied; the first method will preserve the synchrony between the audiovisual signals whereas the second could compromise that harmony.

Since access to a controlled CI users' group for the testing stage is limited, it was vital to look for an alternative perception chain model that may predict CI users' performance under the proposed enhancements. Normal hearing non-native listeners were selected as 'proxy' listeners, due to similarities in perception with CI users including the internal adversity effect, the increased degradation of perception under external noise sources, and the increased sensitivity to visual cues. Auditory training was chosen as a context to test the effect of visual speech enhancement. Studies have shown that exposure to audiovisual speech during auditory training can improve auditory speech perception. Given that, an audiovisual training framework was designed to test the effect of visual speech enhancement on improving the visual speech benefit in training stimuli, and to test the after-training effect on auditory speech perception. A pilot study that evaluated the framework and examined the characteristics of the non-native Saudi group was presented. The observations from this pilot study suggested the effectiveness of the framework in showing the listeners' training gains. Consistent with previous results, the audiovisual training listeners achieved the highest training gain. A gap between native and non-native listeners was found, which could be analogous to the gap between normal hearing native listeners and the CI users [281].

Chapter 5

Appearance Based Enhancement

5.1 Introduction

This chapter presents an appearance based enhancement method, which automatically applies a lipstick effect on a talker's lips in a video to support speech perception. The aim of this effect is to increase the saliency of the visual speech. Evidence from the literature suggests that the saliency of the visual signals can be affected by some facial appearance characteristics that decrease the mouth area's visibility [154, 210, 279] (see Section 4.2). A clear view of the mouth region is crucial to the quality of visual speech, because the lips, teeth, and tongue provide half of the overall visual speech information gathered from the face [210, 297]. In this chapter, the effect of lipstick on the mouth region is addressed by automatically simulating a talker wearing lipstick in audiovisual stimuli. According to Lander and Capeks' [173] observations of talkers who wore real lipstick, this is an effect that can improve lip-reading. The lipstick effect in this chapter is applied directly to the video frames, using image processing techniques to exclude any psychological variables that result from wearing real lipstick.

This effect's impact on the intelligibility of the CI simulated speech using the audiovisual training framework is tested. This is compared with the unaltered-audiovisual and audio-only stimuli. The evaluation study will address two main points:

1. Can the lipstick effect increase the benefits of visual speech for improving the intelligibility of CI simulated speech?

2. Can the lipstick effect increase the audiovisual training gain?

This chapter is organised as follows: Section 5.2 will present a tool that extracts facial features associated with speech; Section 5.3 will present the implementation

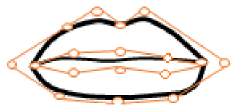
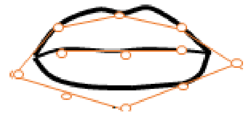
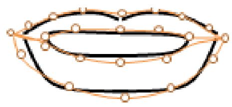
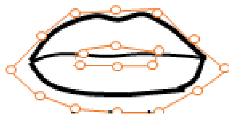
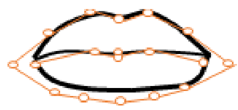
Visage	**Intel Realsense**	**Luxand**
Outer: 10; Inner : 8; Image 	Outer: 12; Inner: 10; Live stream 	Outer: 8; Inner: 6; Image 
FA	**Yu et al.**	**dlib**
Outer:14; Inner: 12; Video 	Outer: 12; Inner: 6; Image 	Outer: 12; Inner: 8; Image 

Table 5.1: A comparison of facial landmark tracking systems. Outer = number of points in the outer mouth contour; Inner = number of points in the inner mouth contour; Processing level = image based or video based. Orange shape: tracked landmarks; Black shape: manual annotation of the mouth

stages of the lipstick effect; and Section 5.4 will present the evaluation of the lipstick effect using the audiovisual training framework (which was introduced in Section 4.4).

5.2 Visual Speech Tracking

The first step in visual speech enhancement is to extract the visual speech information, i.e., facial landmarks associated with the produced speech. An automatic system to track facial features can be used to perform this function. The automation of facial features tracking is an active research area [207, 341]; however, the development of an automatic tracker is considered to be outside the scope of this thesis. As a result, a number of tracking systems and SDKs were tested in order to select a suitable tracker. The tested systems were: *visage* SDK [4]; *Intel RealSense* SDK [3]; *Luxand* SDK [75]; *dlib* toolkit [165] which employs the ensemble regression tree method of Kazemi and Sullivan [157]; Yu *et al.*'s [339] method in localising feature points using a deformable shape model; and *Faceware Analyser* (FA) [2, 262]. Table 5.1 summarises the differences between the tracking systems in terms of the number of lip landmarks and the processing level (image-based or video-based). For each method, the tracked landmarks (in orange) are superimposed on a manual annotation of the lips contour (in black) in the input frame to examine the tracking accuracy. FA was chosen based on the number of landmarks that encode the lip region and the level of accuracy

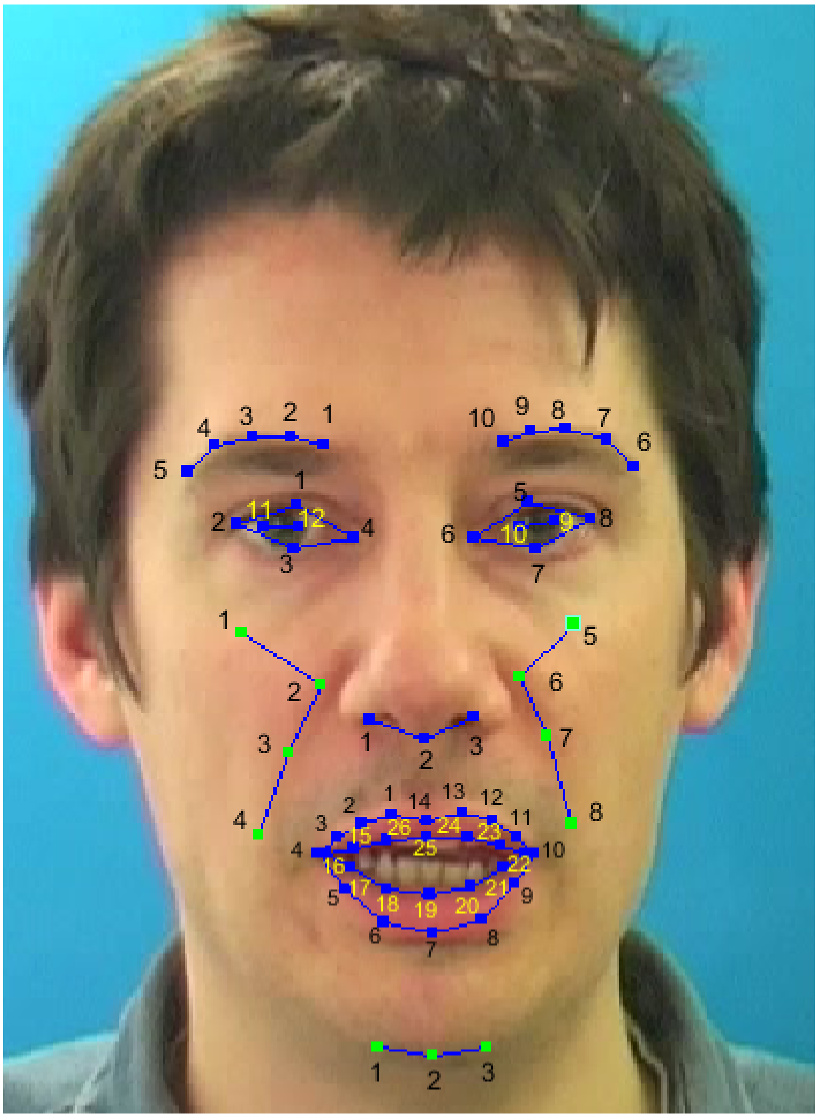


Figure 5.1: The set of facial landmarks extracted by Faceware Analyser [2].

Mouth				**Jaw**		**Eyes**			
						1	eye top right	7	eye bottom left
1	mouth lip outer right	14	mouth lip outer top middle	1	jaw left	2	eye corner outer right	8	eye corner outer left
2	mouth lip outer corner right - 0.3	15	mouth lip inner corner right - 0.6	2	jaw middle	3	eye bottom right	9	pupil outer left
3	mouth lip outer corner right - 0.6	16	mouth lip inner corner right	3	jaw right	4	eye corner inner right	10	pupil inner left
4	mouth lip outer corner right	17	mouth lip inner bottom middle - 0.3	**Cheeks**		5	eye top left	11	pupil outer right
						6	eye corner inner left	12	pupil inner right
5	mouth lip outer bottom middle - 0.3	18	mouth lip inner bottom middle - 0.6	1	cheek upper cheek right	**Brows**			
6	mouth lip outer bottom middle - 0.6	19	mouth lip inner bottom middle	2	cheek nasolabial upper right	1	brow inner right	6	brow outer left - 0.5
7	mouth lip outer bottom middle	20	mouth lip inner corner left - 0.3	3	cheek nasolabial lower right	2	brow middle right - 0.5	7	brow outer left - 0.5
8	mouth lip outer corner left - 0.3	21	mouth lip inner corner left - 0.6	4	cheek lower cheek right	3	brow middle right	8	brow outer left - 0.5
9	mouth lip outer corner left - 0.6	22	mouth lip inner corner left	5	cheek upper cheek left	4	brow outer right - 0.5	9	brow outer left - 0.5
10	mouth lip outer corner left	23	mouth lip inner top middle - 0.3	6	cheek nasolabial upper left	5	brow outer right	10	brow outer left
11	mouth lip outer, left - 0.3	24	mouth lip inner top middle - 0.6	7	cheek nasolabial lower left	**Nose**			
12	mouth lip outer, left - 0.6	25	mouth lip inner top middle	8	cheek lower cheek left	1	nostril outer right	3	nostril outer left
13	mouth lip outer left	26	mouth lip inner corner right - 0.3			2	nose lower middle		

Table 5.2: A description of the FA landmarks.

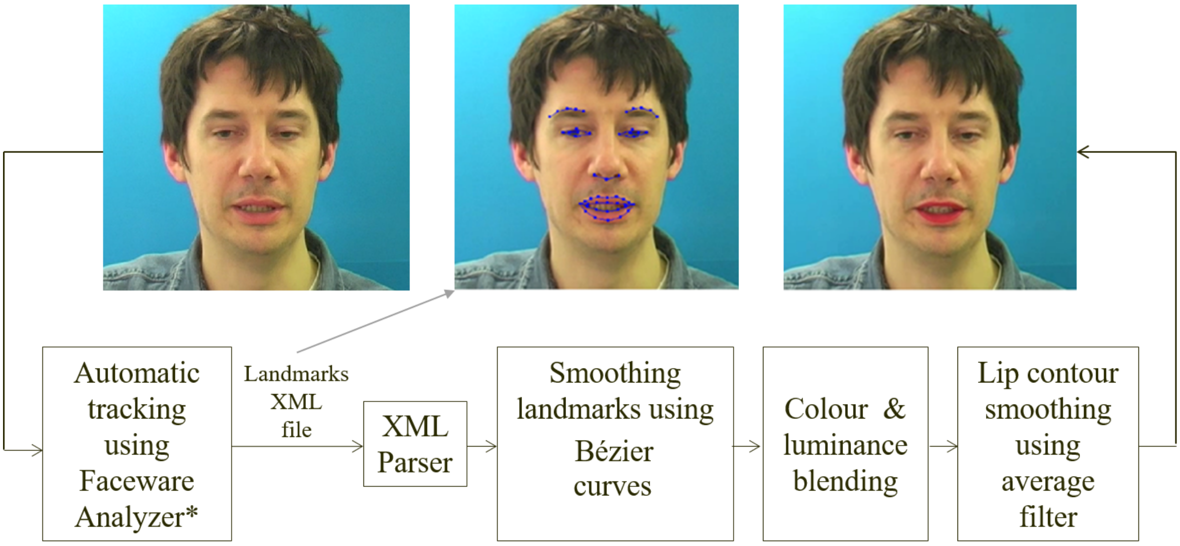


Figure 5.2: Lipstick-effect block diagram.

provided by the software. The tracker module in FA is based on an *Active Appearance Model* (AAM) [58], which uses a statistical model for the shape and the appearance (texture). It is also used to track facial landmarks in a series of images by finding a linear map between changes in shape and texture in a training image set during a training phase. FA applies a *feature locater update model* to the training images, which is derived in steps as follows [262]:

1. A set of training data is defined: in selected video frames (training images), the facial landmarks are annotated;

2. At each frame, a facial features displacement vector, which adjusts the model to the annotated points and the associated texture change vector are generated;

3. A regularised linear regression that maps between the feature displacement vector and the texture change vector is trained.

The features locater model is then used to track facial features beyond the training set. Since AAM may fail to address talker variability as it only produces a generic tracker, FA offers the ability to personalise the tracking for a given talker. This is achieved through a calibration step, in which the optical flow between each video frame and a selected frame of that talker in neutral expression is used to customise the feature locater update model [262].

As a result of the tracking process, FA exports 65 normalised landmarks (lips: 26, jaw: 3, cheeks: 8, nose: 3, eyes: 10, eyebrows: 10) per frame to an XML file.

The individual video frames are also extracted by FA. Figure 5.1 shows an example of a video frame tracked by FA. Table 5.2 provides a description of the landmarks illustrated in Figure 5.1.

5.3 The Lipstick Effect

Figure 5.2 gives an overview of the lipstick effect's implementation. The AV stimuli (Table 4.3) were processed as a batch using FA, generating an XML file and a folder containing JPEG image frames per stimulus. Each XML file was parsed to extract the locations of facial landmarks of interest in all video frames: the 26 landmarks of the mouth included 12 landmarks for the inner mouth region and 14 landmarks for the outer region (see Figure 5.1). The k^{th} video frame may then be associated with two mouth shape vectors:

$$\boldsymbol{in\_lip_k} = \begin{bmatrix} x\_in_1 & y\_in_1 & \cdots & x\_in_{12} & y\_in_{12} \end{bmatrix}^T \tag{5.1}$$

where $\boldsymbol{in\_lip_k}$ is a vector of 24 coordinates of the 12 inner mouth contour points presented in the format (x_in_i, y_in_i) where $1 \leq i \leq 12$, and

$$\boldsymbol{out\_lip_k} = \begin{bmatrix} x\_out_1 & y\_out_1 & \cdots x\_out_{14} & y\_out_{14} \end{bmatrix}^T \tag{5.2}$$

where $\boldsymbol{out\_lip_k}$ is a vector of 28 coordinates of the 14 outer mouth contour points presented in the format (x_out_j, y_out_j) where $1 \leq j \leq 14$.

Smoothing the Mouth Contour A new set of points $\boldsymbol{b\_in\_lip_k}$ and $\boldsymbol{b\_out\_lip_k}$ were generated by fitting Bézier curves to smooth the contours of $\boldsymbol{in\_lip_k}$ and $\boldsymbol{out\_lip_k}$. Bézier curves are parametric curves that approximate the input points (control points). A Bézier curve passes through the first and the last control points, and appears within the convex hull of all control points [329]. A Bézier curve $\boldsymbol{bz}$, which approximates a set of n points $\boldsymbol{p}$ can be expressed as:

$$\boldsymbol{bz}(t) = \sum_{i=1}^{n} b_{i,n}(t)\ \boldsymbol{p_i}, \quad t \in [0,1] \tag{5.3}$$

where $b_{i,n}$ is the Bernstein basis polynomials of degree n expressed as:

$$b_{i,n} = \binom{n}{i} t^i (1-t)^{n-i}, \binom{n}{i} = \frac{n!}{i!(n-i)!}, \quad i = 1 \cdots n \tag{5.4}$$

For the outer lip contours, four Bézier curves are constructed from the lip points (the order of the points is shown in Figure 5.1): a curve to approximate points from

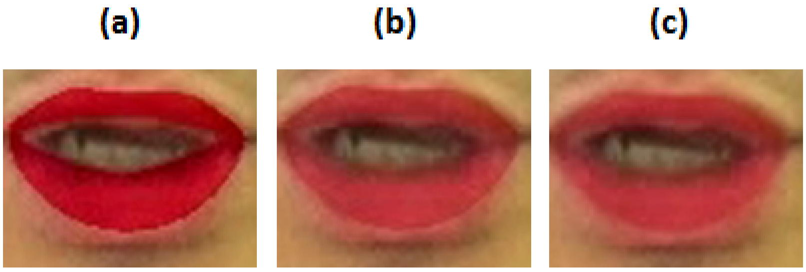


Figure 5.3: (a) Colour blending (b) Luminance blending and (c) The effect of the average filter on smoothing the inner and outer contour after colouring the lips.

1 to 4; a curve to approximate points from 4 to 10; a curve to approximate points from 10 to 13; and a curve to approximate points 13, 14 and 1. Points in all four curves are collectively represented in $\boldsymbol{b\_out\_lip_k}$. For the inner lips, two curve are constructed: one curve for the upper lip points and one curve for the lower lip points. Points in these two curves are represented in $\boldsymbol{b\_in\_lip_k}$.

Colour and Luminance Blending Lip regions bounded by $\boldsymbol{b\_in\_lip_k}$ and $\boldsymbol{b\_out\_lip_k}$ were extracted. Colour blending was applied to each pixel within the area of convex hull of $\boldsymbol{b\_out\_lip_k}$- the convex hull of $\boldsymbol{b\_in\_lip_k}$. Luminance blending was also applied to improve colour blending under different lighting conditions [276]. This was accomplished by identifying the luminance component in each pixel, Y', in Y'CbCr space, adjusting its value in accordance with the image lighting conditions, and then converting the results to the RGB space using the following equations [123]:

$$\begin{bmatrix} Y' \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.1687 & -0.3313 & 0.5 \\ 0.5 & -0.4187 & -0.0813 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{5.5}$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.402 \\ 1 & -0.34414 & -0.71414 \\ 1 & 1.772 & 0 \end{bmatrix} \begin{bmatrix} Y' \\ Cb \\ Cr \end{bmatrix} \tag{5.6}$$

Figure 5.3a and b show the effect of colour and luminance blending. Luminance blending also creates a realistic effect for lipstick, which makes the the surface of the lips appear shiny by clustering the lip regions into a number of intensity clusters based on lip colour (Figure 5.4). Luminance blending is then applied to these clusters in gradual levels — clusters with lighter colours receive a stronger luminance effect and vice versa. The clustering of the lip surface is an iterative process; the clustering stop-criteria is when adding more clusters has a minimal effect on the

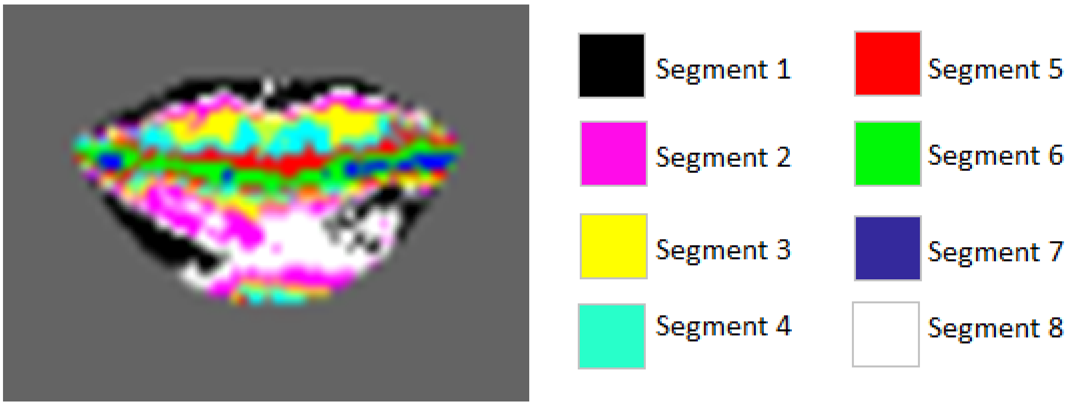


Figure 5.4: segmenting the lips into clusters based on the colour level.

shiny appearance of the lips. Evaluating the minimal effect was based on a subjective decision; Grid videos of the selected talker were grouped based on the lightening conditions during the capture process. Sample videos from each group were taken to identify the ideal number of lip clusters for each group. Figure 5.5a shows a matte (flat) lipstick effect using a uniform intensity level of luminance, whereas Figure 5.5b shows the effect of shiny lipstick using eight luminance levels.

Lip Contour Smoothing After applying colour and luminance blending, the inner and the outer lip regions were stitched back to the talker's face. A 3×3 average filter was also applied to smooth the inner and the outer lips contours (Figure 5.3c). Figure 5.6 shows the lipstick effect on viseme classes extracted from videos of one speaker from the Grid dataset [14]; the first column represents the original viseme shape, while the second column represents the viseme after applying the lipstick effect. This figure shows that applying the lipstick on the mouth region results in more definite mouth shapes and a more prominent appearance of the internal articulators, such as the teeth and the tongue. In the following section, the impact of the lipstick effect on visual speech in audiovisual training is evaluated using the audiovisual training framework.

5.4 Evaluation Study

5.4.1 Audiovisual Training Framework

The audiovisual training framework (Section 4.4) was used to evaluate the effectiveness of the lipstick effect. The framework follows the same methodology of the evaluation study (Section 4.4.1). This includes the use of the Grid corpus

Figure 5.5: (a) Matte lipstick: a uniform level of luminance is applied (b) Shiny lipstick: eight levels of luminance are applied.

to provide the training stimuli, training methodology (an AO pre-test, 3 training sessions, an AO post-test) and baseline training modalities: A training that uses AO stimuli and V training that uses AV stimuli. The audiovisual training framework was adapted to introduce a third training modality, E_1 training, which uses audiovisual stimuli with the lipstick effect (AV_L). AV_L stimuli were produced by applying the lipstick effect to AV stimuli frames. Figure 5.7 and Table 5.3 summarises the training methodology and modalities used to evaluate the effectiveness of the lipstick effect. The recognition scores of subjects attending A and V training will be used to evaluate the effect of the lipstick, i.e., the recognition scores of subjects attending E_1 training. The allocation of subjects into training groups was done automatically by the training tool in the registration stage to ensure randomisation [160].

5.4.2 Subjects

The experiment took place at the female campus of King Saud University in Riyadh, Saudi Arabia. Forty-six non-native Saudi listeners (IELTS score $\in$ [5.5,6]) were recruited in the experiment. All subjects were in the age range 18—40 (Mean = 24 years, SD = 4.5 years). Ethics permission for this study was obtained by following the University of Sheffield Ethics Procedure. Thirteen subjects were assigned to the A training group; 19 subject to the V training group; and 14 subjects to the E_1 training group.

5.4.3 Results

Figure 5.8a shows a comparison of the three subgroups across all training sessions. Table 5.4 summarises these results and Table 5.5 summarises the statistical significance tests results. Generally, introducing the visual signal enhances the mean intelligibility of the vocoded speech (calculated across the three training sessions in Figure 5.8a): the A subjects identified 43% of the AO stimuli, whereas the V subjects

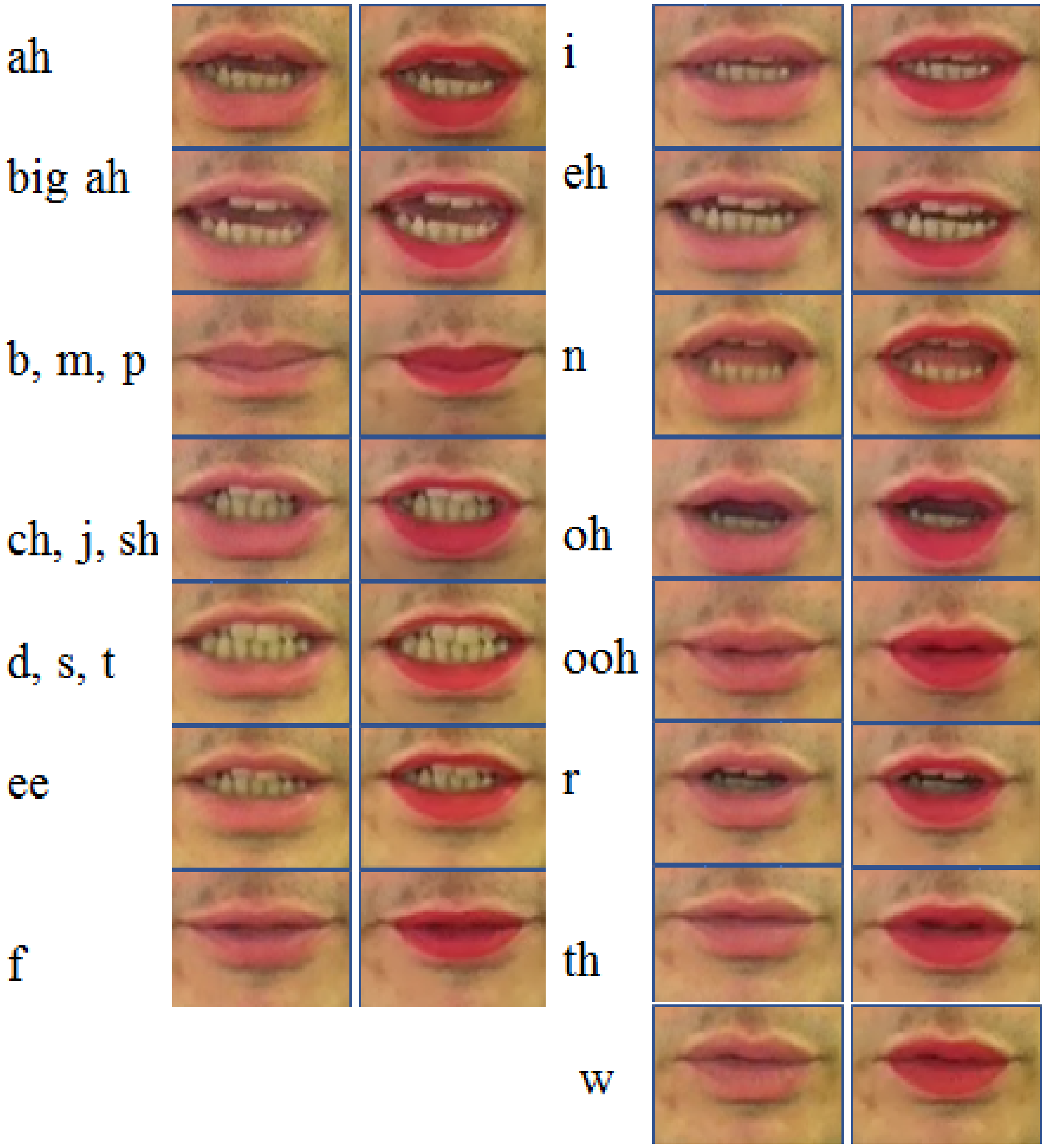


Figure 5.6: Viseme classes extracted for the Grid data sets [14]. The first column represents the original viseme mouth shape while the second column represents the viseme mouth shape after applying the lipstick effect.

	Audio	Video
AO stimuli, A training		-
AV stimuli, V training	CI simulated Grid audio	Grid video
AV_L stimuli, E_1 training		Grid video with lipstick effect

Table 5.3: Stimuli and training modalities.

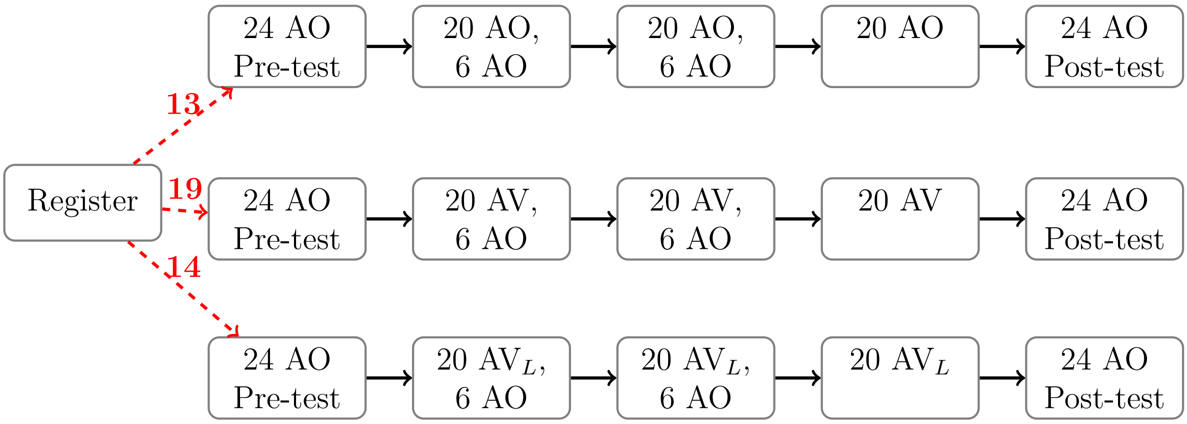


Figure 5.7: The training method consisted of an audio-only pre-test followed by three training sessions then an audio-only post-test session for determining the training gains. The numbers of subjects for each route through the training process are shown in red.

identified 55% of AV stimuli and the E_1 subjects identified 64% of the AV_L stimuli. A one-way ANOVA test showed a significant difference between the mean recognition scores for the three groups, A, V and E_1 ($F(2,43) = 5.36$, $p = .008$) . A post-hoc test, Tukey HSD [315], showed a significant difference between the A and E_1 groups ($p = .006$). No significant difference was found between V and E subjects ($p = .3$), and between the A and V subjects ($p = .1$). The error bars suggest a significant difference between A and (E_1,V) scores in Session 1 scores and E_1 and (A,V) in Session 3 scores. Unlike the other groups, the training profile of the E_1 group features a sharp increase in recognition scores across sessions. Subjects who underwent the E_1 training did not report any problem that can be sourced from the unnaturalness of the lipstick effect; a number of subjects in fact have thought the talker is wearing a real lipstick.

Figure 5.8b shows the mean recognition score that the A, V and E_1 subjects attained in their AO pre- and post-training tests, as well as the mean training gain that subjects achieved in identifying audio-only speech stimuli after they received auditory training. A one-way ANOVA test showed no significant difference between A, V and E_1 subjects in AO pre-test scores ($F(2, 43) = 1.24$, $p = .3$). The scatter plots in Figure 5.9 show all subjects' recognition scores in the AO pre- and post-tests. These plots show good adaptation to the CI simulated speech after training for all subjects, but in variable magnitudes. A one-way ANOVA test showed a significant difference ($F(2, 43) = 3.77$, $p = .03$) between the recognition percentage means in the post-tests of A, V and E_1 subjects. A post-hoc test, Tukey HSD, gave a significant difference between A and E_1 subjects' mean scores in the post-training AO test (p

	A	V	E_1	Vsubset	E_1subset
Pre-test	14%	19%	25%	14%	13%
Post-test	46%	56%	70%	54%	71%
Training gain	32%	37%	50%	40%	58%

Table 5.4: The AO pre- and post-test mean recognition score and the training gain for A, V and E_1, Vsubset, and E_1subset subjects.

= .02). No significant difference was found between the V and E_1 subjects (p = .2), and between A and V subjects (p = .4). Figure 5.8c shows a significant difference in E_1 subject recognition scores between milestone 1 and 2.

The comparison between V and E_1 subjects' scores suggests an increase in the benefit of visual speech on improving speech intelligibility during and after the training when using the lipstick effect. It is a valid comparison between the two subgroups given their comparable training starting points (session 1 scores – Figure 5.8a) and their pre-test abilities (Figure 5.8b).

However, A subjects attained the lowest pre-test recognition mean score compared with the V and E subjects, which complicates the comparison of the effect of the training modality on the training gain between the three subgroups. A possible solution is to use subsets of the E_1 and V subjects that match the pre-test ability of the A subjects as closely as possible. This can be done by sorting the list of subjects from the corresponding group based on the pre-test results, and then removing those of highest ability such that the pre-test abilities of the remaining subjects in the list are equivalent to the A subjects' pre-test ability. Using Figure 5.9, V subject numbers 15 and 16 were removed to create Vsubset, and E_1 subject numbers 1, 2, 9, 10, 12 and 13 were removed to create E_1subset.

Figure 5.10b shows the mean recognition scores for the pre and post-tests for A, Vsubset and E_1subset subjects. Tables 5.4 and 5.5 summarise these results. A one-way ANOVA test shows a significant difference in post-test mean recognition percentage for A, Vsubset and E_1subset subjects ($F(2,35) = 3.32$, $p = .04$) and in training gain ($F(2,35) = 5.04$, $p = .01$). A post-hoc test, Tukey HSD, gave a significant difference between A and E_1subset subjects mean recognition percentage in the post-training AO test ($p = .037$) and in training gain ($p = .009$). No significant difference was found between Vsubset and E_1subset subjects in mean post-test recognition scores ($p = .23$) and training gain ($p = .085$). Also, no significant difference was found between A and Vsubset subjects in mean recognition percentage ($p = .5$) and training gain ($p = .4$).

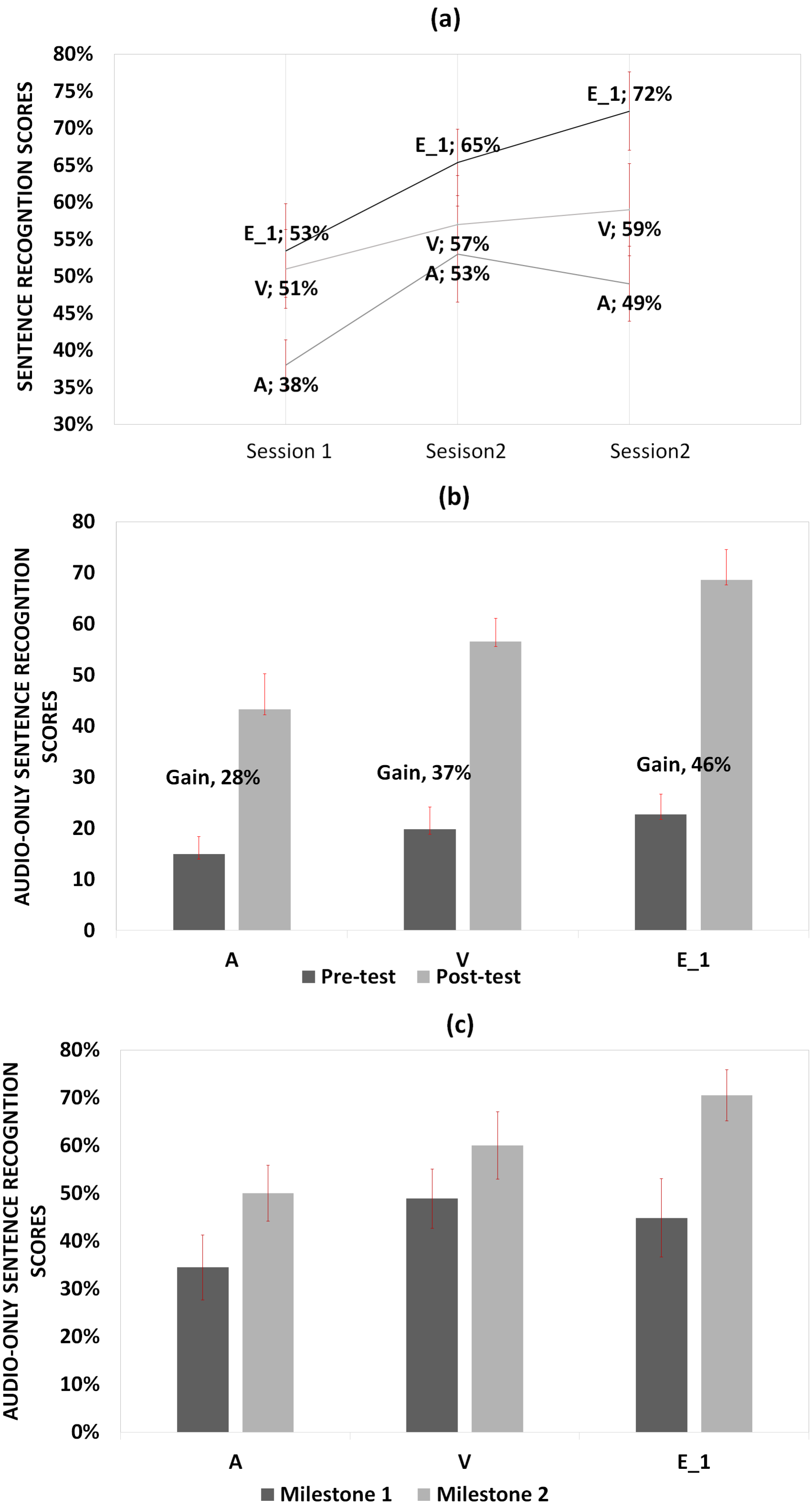


Figure 5.8: All data results for the A, V, E_1 subjects:(a) sentence recognition during training; (b) audio-only pre- and post-test mean recognition scores and training gains $(posttest - pretest)$; (c) training impact on audio-only sentence recognition (learning milestones). Errors bars =/- standard error.

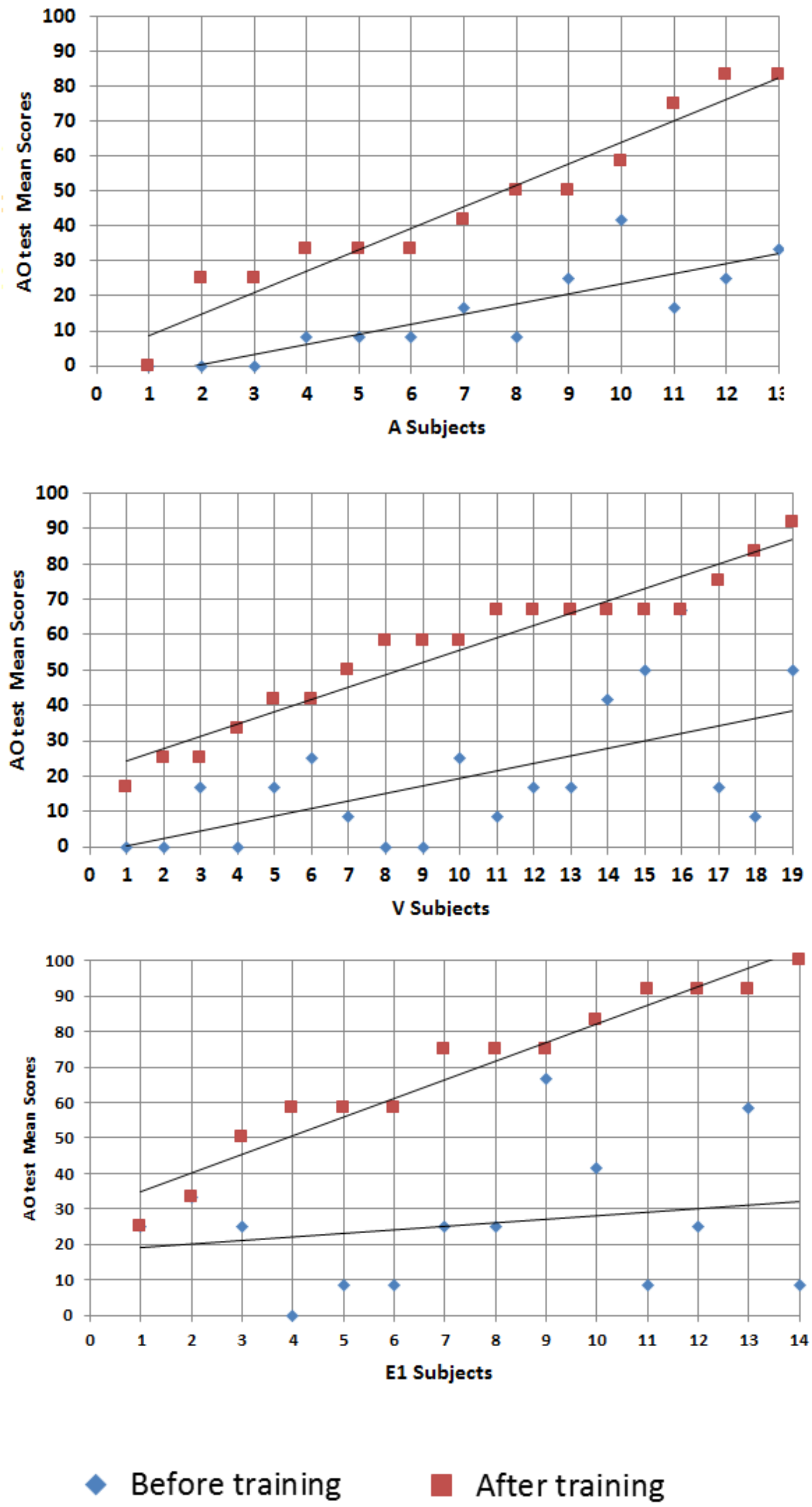


Figure 5.9: The mean recognition scores for the AO tests before and after training for A, V and E_1 subjects.

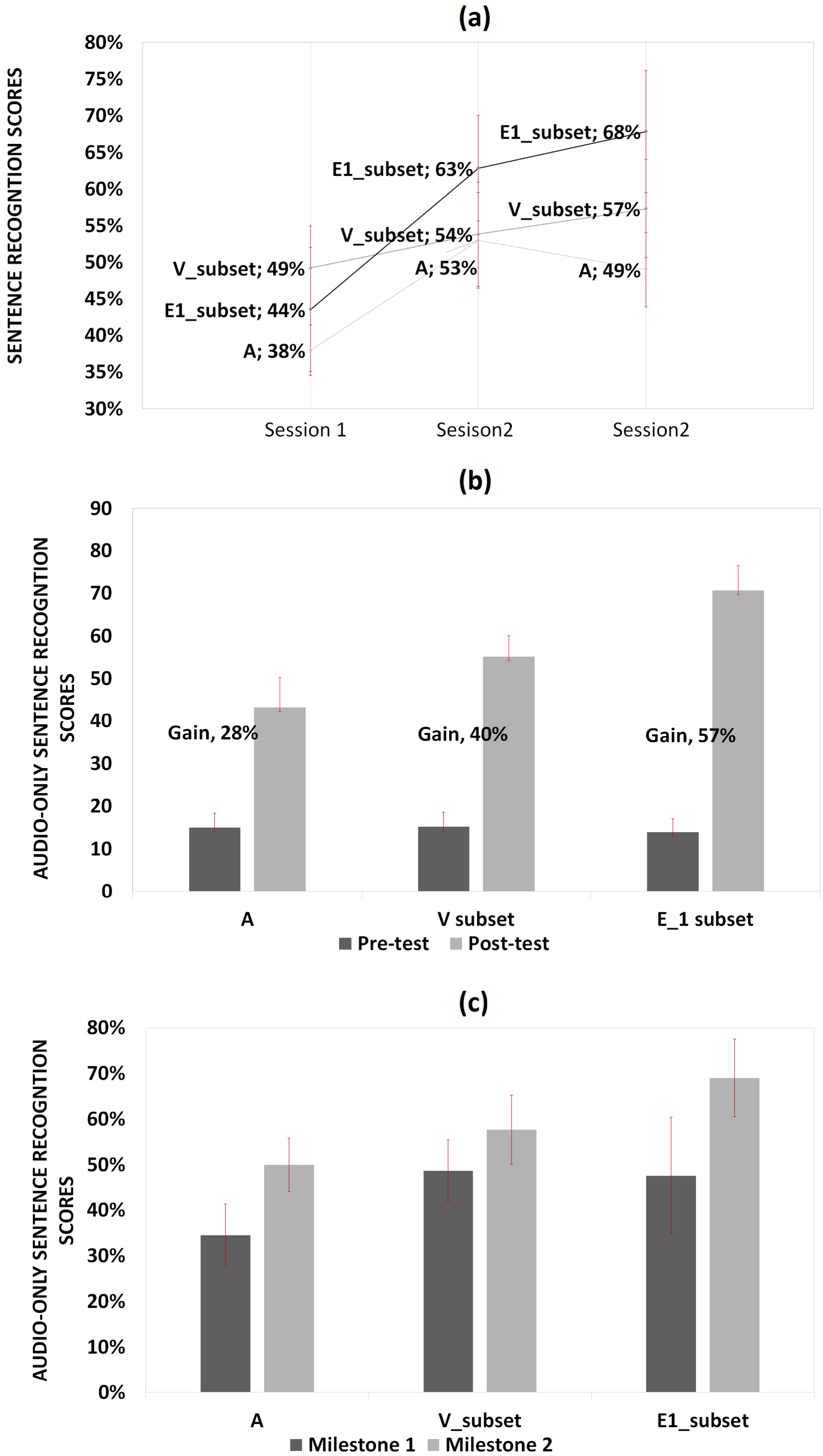


Figure 5.10: Subset results for the A, V, E_1 subjects:(a) sentence recognition during training; (b) audio-only pre- and post-test mean recognition scores and training gains ($posttest - pretest$); (c) training impact on audio-only sentence recognition (learning milestones). Errors bars =/- standard error.

Sentence recognition during training		
V > A	**E_1 > A**	**E_1 > V**
0.1	0.006	0.3
Vsubset > A	**E_1subset > A**	**E_1subset > Vsubset**
0.1	0.007	0.2
Post-test mean recognition scores		
V > A	**E_1 > A**	**E_1 > V**
0.4	0.02	0.2
Vsubset > A	**E_1subset > A**	**E_1subset > Vsubset**
0.5	0.03	0.23

Table 5.5: Summary of the results of statistical analyses (p-value) between two training modalities. Symbols: > mean scores by first training's subjects is garter than mean scores by second training's subjects.

Vsubset and E_1subset showed comparable results to V and E_1 subjects in speech intelligibility scores (Figure 5.10a). Overall, Vsubset subjects identified 55% of audiovisual speech stimuli whereas E_1subset subjects identified 67% of the enhanced audiovisual stimuli. A one-way ANOVA test showed a significant difference between the recognition mean percentage for A, Vsubset and E_1subset subjects ($F(2,35) = 5.45$, $p = .009$). A post-hoc test, Tukey HSD, showed a significant difference between the intelligibility scores of the A and E_1subset subjects ($p = .007$). No significant difference was found between the Vsubset and E_1subset subjects ($p = .2$), and between A and Vsubset ($p = .1$). Error bars suggest no significant difference between all groups across sessions. Figure 5.10c confirms the same finding as shown in Figure 5.8c which found a significant difference in E_1 subject recognition scores between milestone 1 and 2.

Confusion matrices (Figure 5.11) were produced to understand the possible sources of confusion the subjects had while identifying the letter keywords during the training and AO post test. Letters that were not uniformly presented across the subgroups were omitted from the matrices. The E_1 subjects (mean recognition score: training: 70%; testing: 75%) were less confused than the V (mean recognition score: training: 60%; testing: 65%) and A subjects (mean recognition score: training: 50%; testing: 55%). A one-way ANOVA test between A, V and E_1 subjects' mean letter recognition scores in AO post-test confirmed a significant difference ($F(2,63) = 4.18$, $p = .019$). The post-hoc test, Tukey HSD, showed a difference between the A and E_1 subjects ($p = .01$). No significant difference was found between the V and E_1 subjects ($p = .2$), between the A and V subjects ($p = .4$), and between the A, V and E_1 subjects in letter recognition scores in the training.

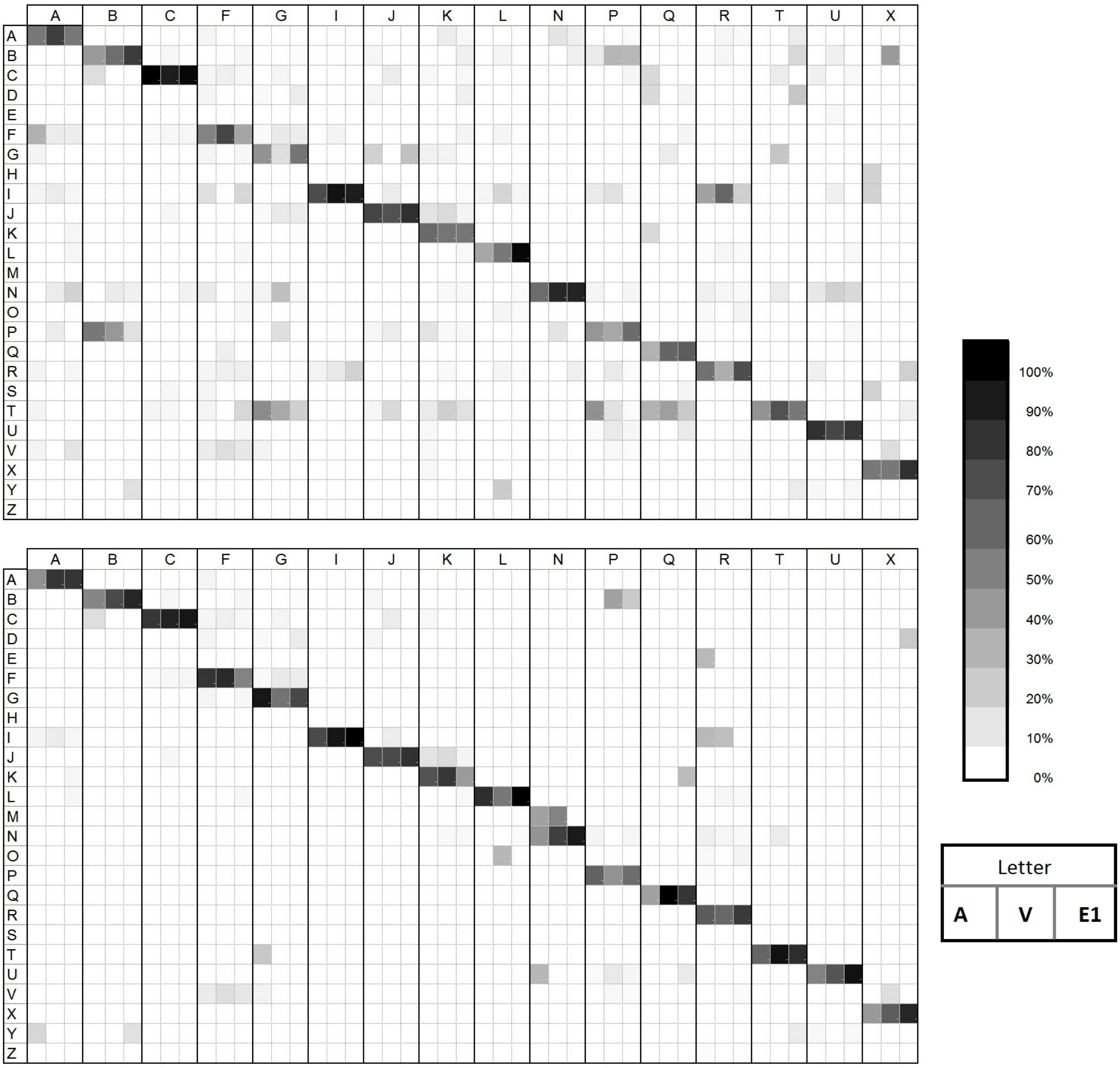


Figure 5.11: Confusion matrix of letter-recognition scores during (top) training and audio-only post-testing (bottom) for A, V, E_1 subjects. Each cell is divided into 3 sub-cells: A, V, and E_1 (from left to right). Colour shades represents the scale of recognition/confusion mean rate (the darker the shade, the higher the response to this cell).

Across the letters, introducing the visual signal helped the V subjects to achieve higher scores (63% and 66% in training and test, respectively) in the recognition of letters that are constructed from diphthongs [a, e, i, u] compared with A subjects (53% and 52%). This is also true for E_1 subjects (72% and 82.5%). A significant difference was found between E_1 and A subjects in diphthong recognition ($F(2,14) = 5.16$, $p = .01$). No significant difference was found between the V and E_1 subjects ($p = .2$), and between the A and V subjects ($p = .2$). Also, no significant difference was found in identifying letters with plosive or fricative sounds across the subjects.

Introducing the visual signal in auditory training was found to impede learning of visually similar letters such as G and T, and P and B in the study reported in Section 4.4. This effect was also examined in this study. V subjects showed higher confusion rate in identifying P (confusion rate = 60%) and G (confusion rate = 48%). E subjects were slightly less confused than V subjects in identifying P (confusion rate = 46%) and G (confusion rate = 32%). This may indicate that the enhanced visual cues become more salient and distinguishable for the E_1 subjects.

5.4.4 Discussion

This study had a two-fold aim: to investigate the usefulness of applying an enhancement to visual speech used in audiovisual training; and to evaluate the effect of the lipstick enhancement. The main observation and discussion points are as follows.

The Impact of Visual Speech Using Sumby and Pollacks' [296] metric, the visual contribution to speech recognition under noisy conditions can be quantified as follows:

$$C_V = \frac{C_{AV} - C_{AO}}{1 - C_{AO}} \tag{5.7}$$

where C_{AO} and C_{AV} are the normalised recognition scores of AO and AV, respectively. Similarly, the visual contribution of the lipstick effect can be expressed as follows:

$$C_{VL} = \frac{C_{AV_L} - C_{AO}}{1 - C_{AO}} \tag{5.8}$$

where C_{AV_L} is the normalised recognition score of AV_L. C_V and C_{VL} can be calculated from the intelligibility scores in the last training session (Figure 5.8a and 5.10a). They show that applying the lipstick effect has increased the visual speech contribution in enhancing the intelligibility of CI simulated speech from 0.19 to 0.45 in all data, and 0.15 to 0.37 in the subset data. The results here confirm Lander and Capeks' [173]

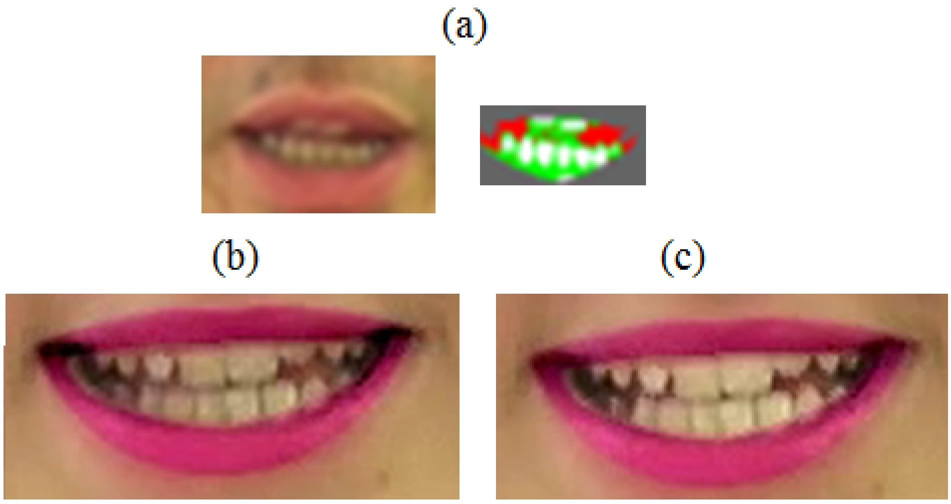


Figure 5.12: Internal and external articulator enhancement: applying the lipstick effect in conjunction with increasing the luminance blending of the teeth and the tongue. (a) teeth extraction by thresholding then segmentation; (b) lipstick only; (c) lipstick and teeth effect.

findings on the effect of lipstick on supporting speech-reading. They argued that the effect of the lipstick might be from the talkers making exaggerated mouth movements. However, the findings in this chapter confirm the effect of the lipstick on improving speech intelligibility.

The Impact on the Training Gain Consistent with previous findings [28, 250], and the results from Section 4.4.1, introducing visual speech in auditory training (i.e., V and E_1) has increased the training gain. E_1 and E_1-subset subjects, however, achieved the highest training gain amongst all subgroups, with significant difference from the A group, and also a significant improvement in letter recognition compared with the other subsets. Such findings suggest stronger auditory discrimination abilities for the E_1 and E_1-subset subjects. This may mean that enhancing the visual speech has helped to improve the role of visual signals in guiding the perceptual learning of auditory-only skills. The training milestones results also support this as they show a fast learning effect in the E_1 and E_1-subset subjects.

Increased Attention or Increased Visual saliency? Lander and Capek argued that the lipstick's effect may have resulted from increased 'selective attention' towards the talker's lips and suggested using an eye-tracking method to measure the time that listeners spent on looking at the mouth area. However, humans are reportedly unable to maintain selective attention on a task for more than 20 minutes [59]. The training in this study took on average one hour per subject and the E_1 group's training profile showed a steep improvement, reaching its maximum at the last training session. Given the short attention span in humans and the long training duration, the increase in the visual speech benefit on speech intelligibility may have resulted from an increased visual speech saliency from applying the lipstick effect, rather than an increase in selective attention towards the talker's mouth. Another observation that may support the hypothesis of increased visual saliency is that visually similar phoneme pairs with high confusion rates for the un-enhanced visual speech became less confused when the lipstick effect was applied. This suggests that the lipstick effect may have offered extra cues that helped the listeners to discriminate between pairs that are intrinsically similar.

The Usefulness of Visual Speech Enhancement The impact of the lipstick effect on increasing the benefit of visual speech and the training gain suggests the great potential of visual speech enhancement that preserves synchrony between the audio and the visual speech signals. These enhancements could be applied to enhance audiovisual speech intelligibility on different platforms such as TV and YouTube. Future work could investigate the effect of increasing the saliency of the internal articulators such as the teeth and the tongue (Figure 5.12).

5.5 Summary

This chapter presented the implementation and evaluation of the appearance based visual speech enhancement — the automatic lipstick effect. First, FA presented as a tool for extracting facial features associated with visual speech. The implementation stages for the lipstick effect were then introduced. The impact of the lipstick effect was evaluated using the audiovisual training framework. The evaluation produced two sets of results: data from all subjects, and data from selected subjects who showed comparable pre-test abilities. Both sets of results confirmed the lipstick effect's positive impact on increasing the benefit of visual speech on speech intelligibility, and on improving the training gain of audiovisual training. These findings support the usefulness of the visual speech enhancement in audiovisual training and encourage the investigation of more enhancement methods. In the next chapter, the implementation and evaluation of the impact of a kinematics based enhancement (an exaggeration effect) will be presented.

Chapter 6

Kinematics Based Enhancement

6.1 Introduction

Speaking style is a determining factor in the quality of visual speech [154, 279]. According to Lindblom's hypo-hyper (H&H) theory of speech production [187], talkers make articulatory energy modifications from hypo- to hyper-articulated speech in order to adapt to the demands of the listening situation. This may create a variety of speaking styles that exert different energy magnitudes in order to move the external articulators (Figure 6.1) [83]. H&H theory provided the motivation to investigate the transition from hypo- to hyper-articulated speech as a method for enhancing visual speech.

Given a video that features a talker speaking normally, the enhancement method presented in this chapter depends solely on the visual speech kinematics data presented in the input video, with no a priori knowledge about the hyper-articulation style of the talker. This kinematics enhancement method uses, with modification, Theobald *et al.'s* [309] exaggeration method that amplifies a talker's mouth shapes and appearance to produce a more pronounced speaking style. Theobald *et al.* tested the visual perception of the exaggerated visual stimuli and found improved lip-reading performance among inexperienced lip-readers. In this chapter, the effect of visually exaggerated audiovisual stimuli on audiovisual perception is examined.

Whilst the kinematics based enhancement approach increases the saliency of visual speech, it may compromise the harmony between the audio and visual aspects in speech. This is because in the kinematics based enhancement, the exaggeration effect on the audiovisual stimuli is only applied to the visual speech signal, whereas the audio speech signal remains un-exaggerated.

This chapter will address the following questions:

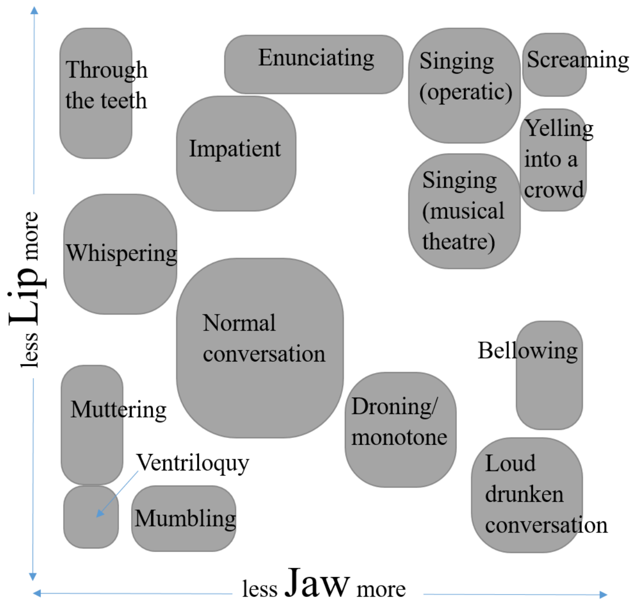


Figure 6.1: The transition from hypo- to hyper-articulation in the Jaw Lip viseme model (JALI), which simulates visual speech by aggregating functions related to jaw motion and lip motion [83]. This shows that distinct speaking styles exert different articulatory-energy modifications. (Permission to use this image has been granted by Edwards [83].)

- Can listeners adapt to the conflict between the audio speech and the exaggerated visual speech when they are presented together?
- Can the exaggeration effect increase the CI simulated speech intelligibility?
- Can the exaggeration effect increase audiovisual training gain when applied to audiovisual training stimuli?

Section 6.2 will present the implementation of the exaggeration effect. Section 6.3 will present the evaluation that investigates the impact of the exaggeration effect on the visual benefit of audiovisual speech and on the training gain when using the audiovisual training framework.

6.2 The Exaggeration Effect

6.2.1 Modeling Exaggerated Mouth Shapes

Principal Component Analysis (PCA) is a nonparametric technique to reduce the dimensionality of input data by finding the orthogonal axes of variation using *eigenvectors* and their *eigenvalues*. The eigenvectors and the eigenvalues are generated from the *covariance matrix* of the data; the eigenvectors correspond to the directions of the most variance, whereas the eigenvalues quantify the variance for the associated eigenvectors. Principal components (PCs) can then be described as a weighted sum of variables (eigenvectors) given their contributions in making the variance in the data [137].

Modelling the exaggeration of the mouth shapes in a given video can be achieved by extrapolating the mouth shapes' Principal Components (PCs). Eigenvectors can be considered as the main building blocks, or gestures, that make up mouth shapes. Such gestures have variable contributions in making each mouth shape for each video frame, which explains the variation in the mouth shape data. Describing a mouth shape in terms of its basis gestures and their contributions can offer a shape parametrisation that can be used to guide a controlled extrapolation of the mouth shapes.

Using FA (Section 5.2), mouth shape data (26 landmarks of the inner and outer lips) were extracted from AV stimuli along with other face data (30 landmarks of eyebrows, eye corners, pupil, and nose) – see Figure 5.1. To correct for small variations over time in the talker-camera distance in a video, the mouth coordinates were normalised and translated to produce a zero-centred mouth space prior to applying PCA. The points were normalised by dividing the mouth landmarks in a frame k by $\boldsymbol{d_k}$, where d_k is the Euclidean distance between the midpoint of the inner corners of the eyes and the tip of the nose, since these are assumed to be unaffected by the articulation. To create the zero-centred mouth model space, the normalised mouth landmarks in frame k were translated by $\boldsymbol{t}$ to be aligned with the centre of the normalised mouth landmarks in the first frame, where t is formed from the 2D distance between the mouth centres. The k^{th} video frame can then be associated with two vectors: a mouth shape vector of 52 elements, expressed as:

$$\boldsymbol{lip_k} = \begin{bmatrix} x\_l_1 & y\_l_1 & \cdots & x\_l_{26} & y\_l_{26} \end{bmatrix}^T \tag{6.1}$$

and a face shape vector of 60 elements, expressed as:

$$\boldsymbol{face_k} = \begin{bmatrix} x\_f_1 & y\_f_1 & \cdots & x\_f_{30} & y\_f_{30} \end{bmatrix}^T \tag{6.2}$$

The set of eigenvectors generated by a covariance matrix of a given training set can be used to approximate any of that set [57, 309]. A set of eigenvectors can be generated by the covariance matrix $\boldsymbol{C}$ of mouth shapes from a given video

$(\boldsymbol{lip_m}, \boldsymbol{lip_{m+1}}, \cdots \boldsymbol{lip_{m+n}})$, where n = all video frames - silence frames[1], and m is the index of the first non-silence frame in the input video sequence. $\boldsymbol{C}$ is defined as:

$$\boldsymbol{C} = \frac{1}{n}\sum_{k=m}^{n+m}(\boldsymbol{lip_k} - \overline{\boldsymbol{lip}})^T(\boldsymbol{lip_k} - \overline{\boldsymbol{lip}}) \tag{6.3}$$

where $\overline{\boldsymbol{lip}}$ is the mean mouth shape in the corresponding video, defined as

$$\overline{\boldsymbol{lip}} = \frac{\sum_{k=m}^{n+m}\boldsymbol{lip_k}}{n} \tag{6.4}$$

A weighted sum of the eigenvectors is then used to approximate any mouth shape, $\boldsymbol{lip_k}$, in that video as follows:

$$\boldsymbol{lip_k} \approx \overline{\boldsymbol{lip}} + \boldsymbol{P}\,\boldsymbol{b_k} \tag{6.5}$$

where $\boldsymbol{P}$ is the matrix of h eigenvectors (basis gestures) with the highest eigenvalues. h is the number of basis gestures that can account for 90-99% of the lip variance; and h $\approx$ 5 based on tests made on selected videos. $\boldsymbol{P}$ (where each column represents a basis gesture), can be expressed as:

$$\boldsymbol{P} = \begin{bmatrix} p_{1,1} & p_{2,1} & \cdots & p_{x,1} \\ p_{1,2} & p_{2,2} & \cdots & p_{x,2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1,52} & p_{2,52} & \cdots & p_{x,52} \end{bmatrix} \tag{6.6}$$

and $\boldsymbol{b_k}$ is an h-dimensional vector, expressed as:

$$\boldsymbol{b_k} = \begin{bmatrix} b_{k,1} & b_{k,2} & \cdots & b_{k,h} \end{bmatrix}^T \tag{6.7}$$

and given by:

$$\boldsymbol{b_k} = \boldsymbol{P^T}(\boldsymbol{lip_k} - \overline{\boldsymbol{lip}}) \tag{6.8}$$

where $\boldsymbol{b_k}$ defines the contribution of each basis gesture in the representation of $\boldsymbol{lip_k}$, which can be seen as a measure of the distance between $\overline{\boldsymbol{lip}}$ and $\boldsymbol{lip_k}$ [309]. Figure 6.2 shows the first five modes of variation of a talker's mouth shape in a selected AV stimulus (ID = bwwa2p). Each mode of variation was constructed by adding a weighted basis gesture (i.e., multiplied by its contribution) to the mean shape. Varying the contribution of the basis gesture consequently modifies the mouth shape. The first and the second modes represent the increase/decrease in the vertical and the horizontal mouth aperture, respectively. The third and fourth modes address

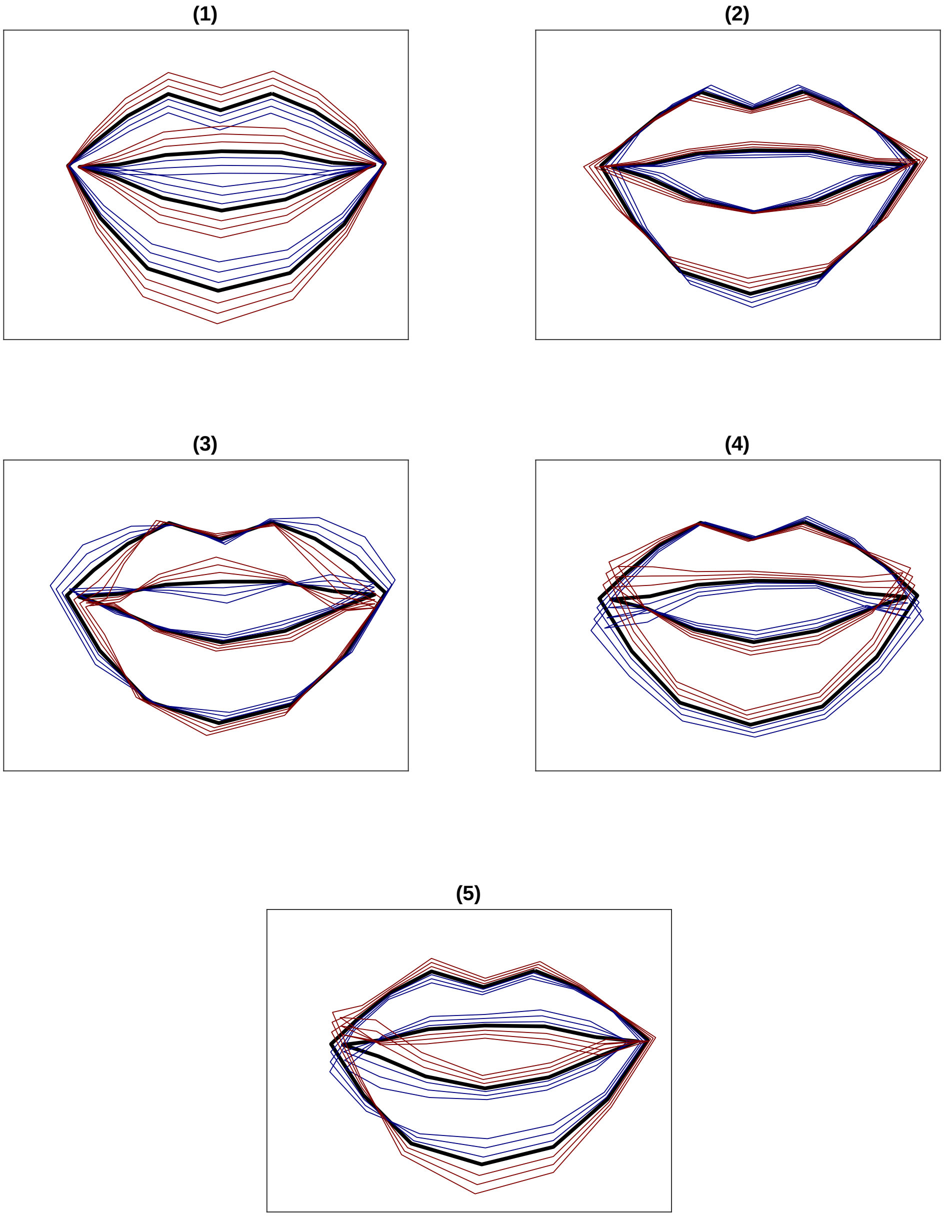


Figure 6.2: The first five modes for a mouth shape model in a selected AV stimulus (ID= bwwa2p), each constructed from a basis mouth gesture. The change in the mouth shape size corresponds to the change in the basis gesture contribution: changes from the mean by +(blue)/-(red) 3 standard deviations.

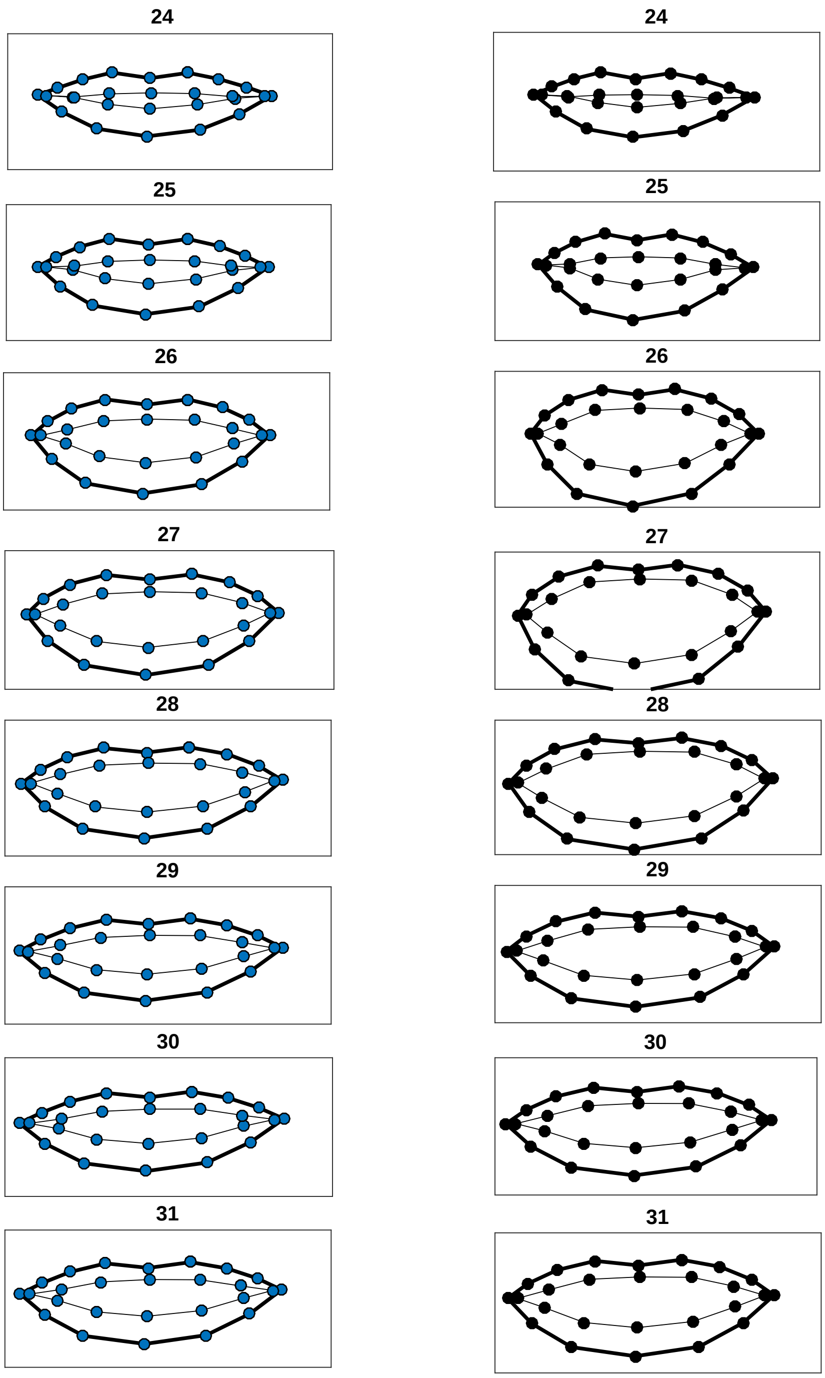


Figure 6.3: The exaggeration effect on frames 24-31 selected from AV stimulus ID = bwaa2p. Left column: plain shapes; right column: exaggerated shapes.

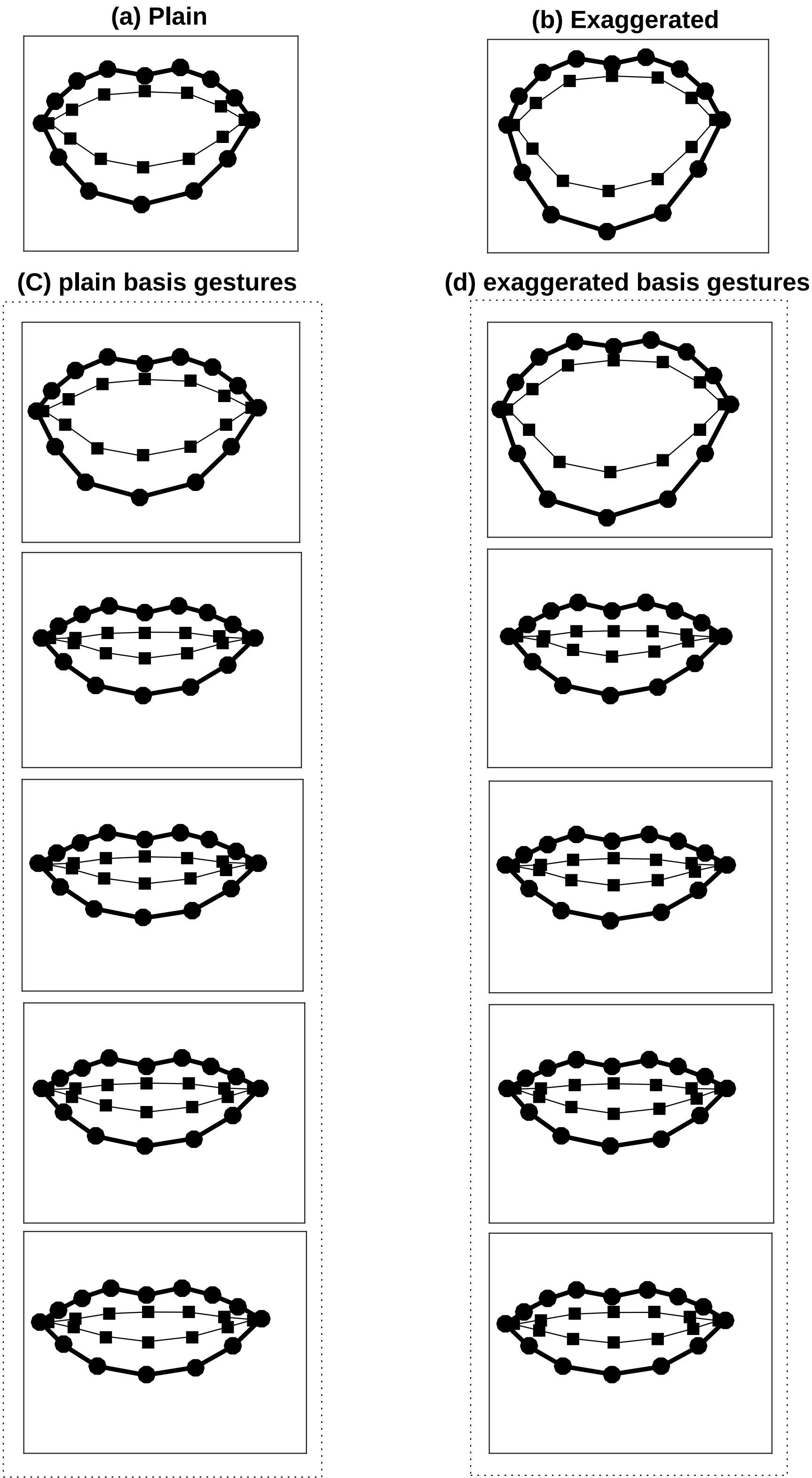


Figure 6.4: Frame 28: (a) Plain and (b) Exaggerated mouth shapes and (c, d) their basis gestures, respectively.

the increase/decrease in the lower and the upper lip movement, respectively. The fifth mode represents the increase/decrease in mouth rounding.

Multiplying $\boldsymbol{b_k}$ with a scalar, $para > 1$, can extrapolate (i.e., exaggerate) lip shape as follows:

$$\boldsymbol{newlip_k} \approx \overline{\boldsymbol{lip}} + para\, \boldsymbol{P}\, \boldsymbol{b_k} \tag{6.9}$$

To project $\boldsymbol{newlip_k}$ from the zero-centred model space back to its location in the video frame, $\boldsymbol{newlip_k}$ is translated by $\boldsymbol{t^{-1}}$ and then scaled by $\boldsymbol{d_k}$. Figure 6.3 shows the impact of exaggeration on selected frames from an AV stimulus (ID = bwwa2p). Figure 6.4 shows plain and exaggerated basis gestures of frame 28 in the selected AV stimulus.

Frames were re-animated by applying a 2D piecewise linear warping method using the estimated exaggerated mouth shapes $\boldsymbol{newlip_k}$ to apply the exaggeration effect. The following section outlines the image warping process.

6.2.2 Image Warping

Image warping is applied in order to produce the exaggeration effect on the video frames. Image warping is a transformation that maps points from one plane to another. It can be either parametric such as translation, bilinear and polynomial transformation, or non-parametric such as piecewise affine transformations [115]. A piecewise affine transformation approach was favoured in this work since it shows better performance in considering local distortions than the parametric methods. The main steps that constitute the piecewise image warping method are:

1. The selection of control points that represent the image points before and after the target transformation is applied. The selected control points are characterised as being the centre of the gravity in the image points [118].

2. The convex hulls of the selected control points are partitioned into triangles using a triangulation algorithm.

3. A correspondence (mapping function) between a source triangle in convex hull 1 and a destination triangle in convex hull 2 is inferred and used to guide the interpolation of the destination triangle pixels in the image.

[1] n$\approx$ 51; the number of frames in Grid videos $\approx$ 64 frames and the average of silence frames $\approx$ 13 frames

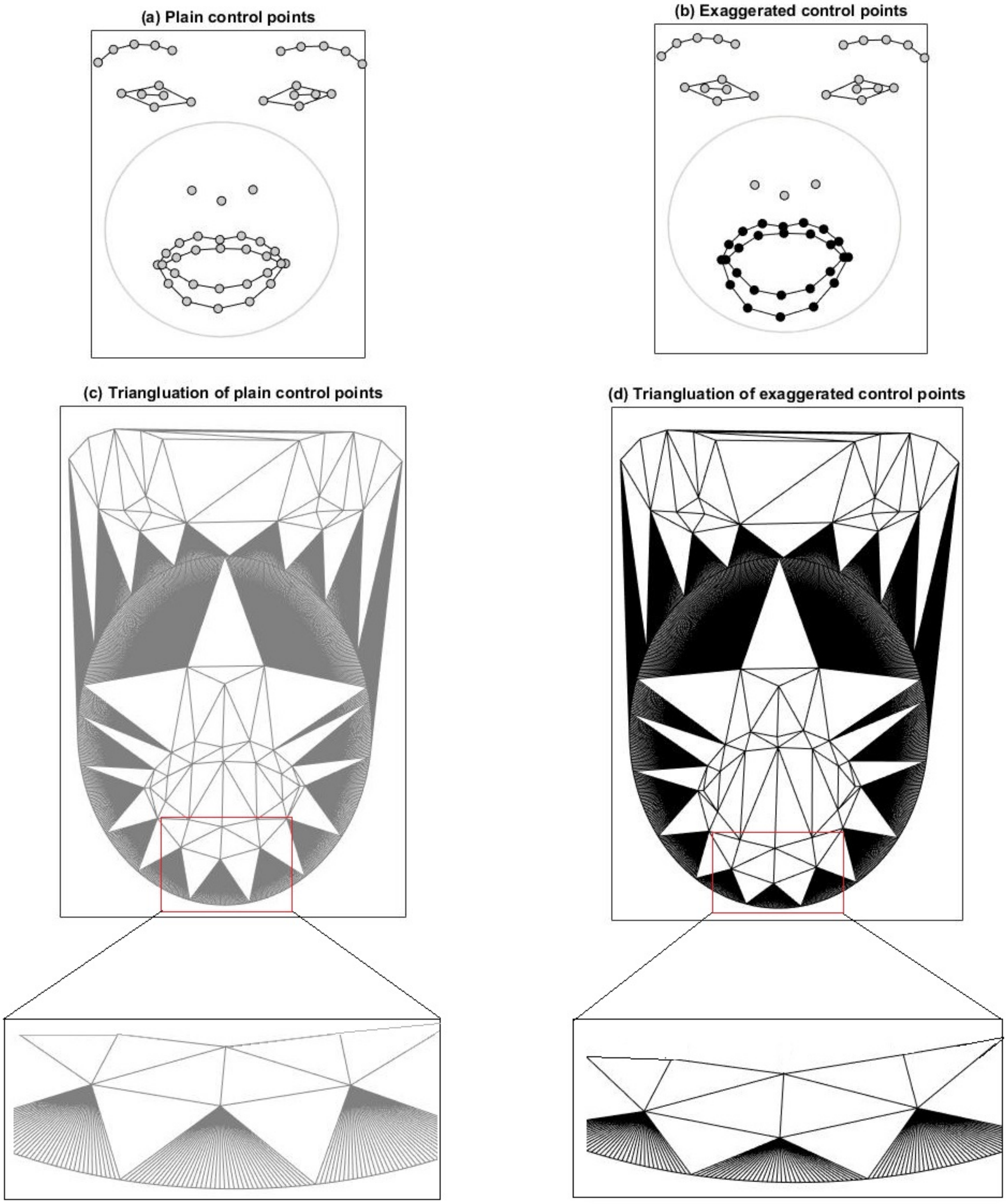


Figure 6.5: Triangulation of the control points in Frame 28.

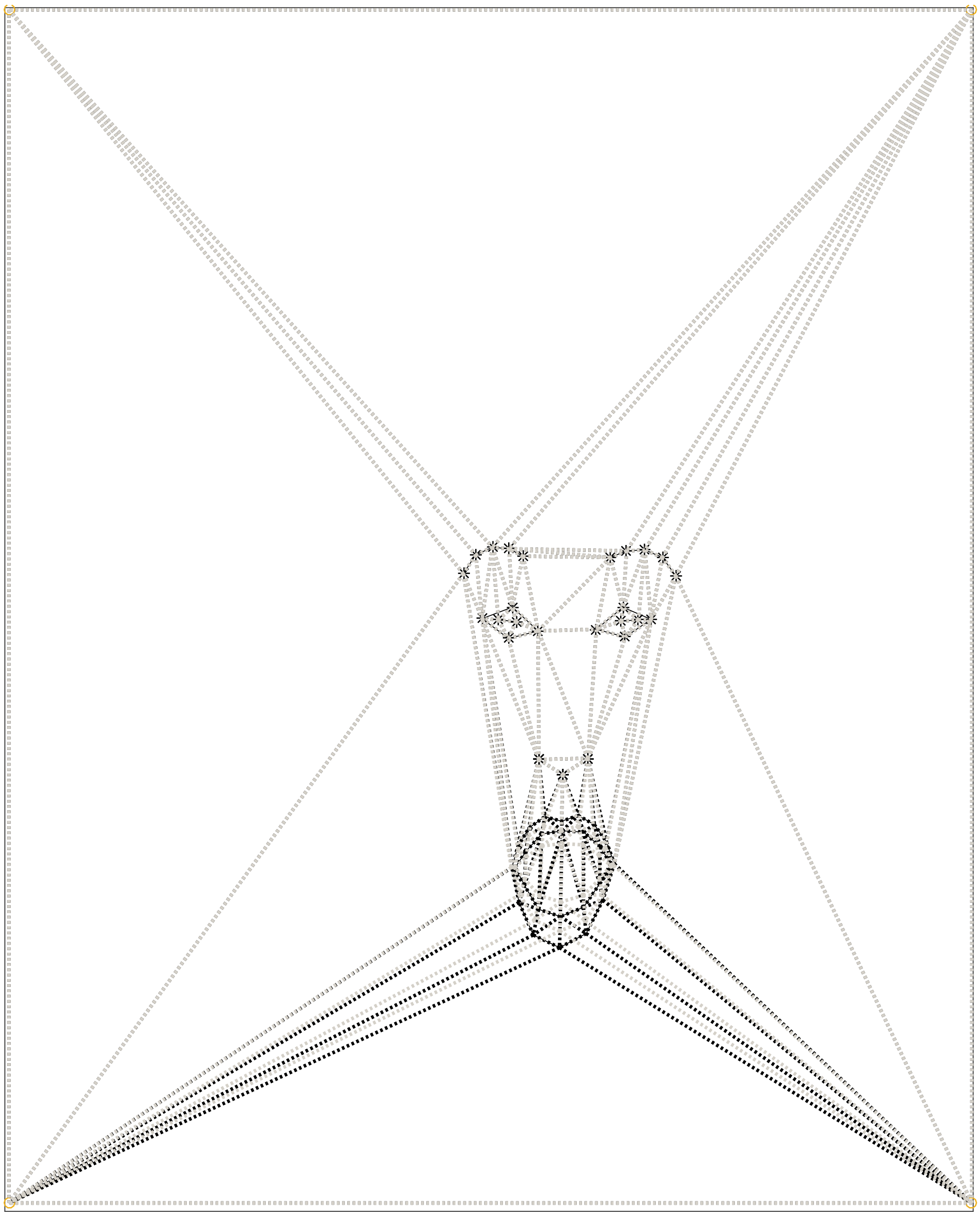


Figure 6.6: Triangulation of the control points of frame 28 without using the rings; Gray: ***basline***, Black: ***exag***.

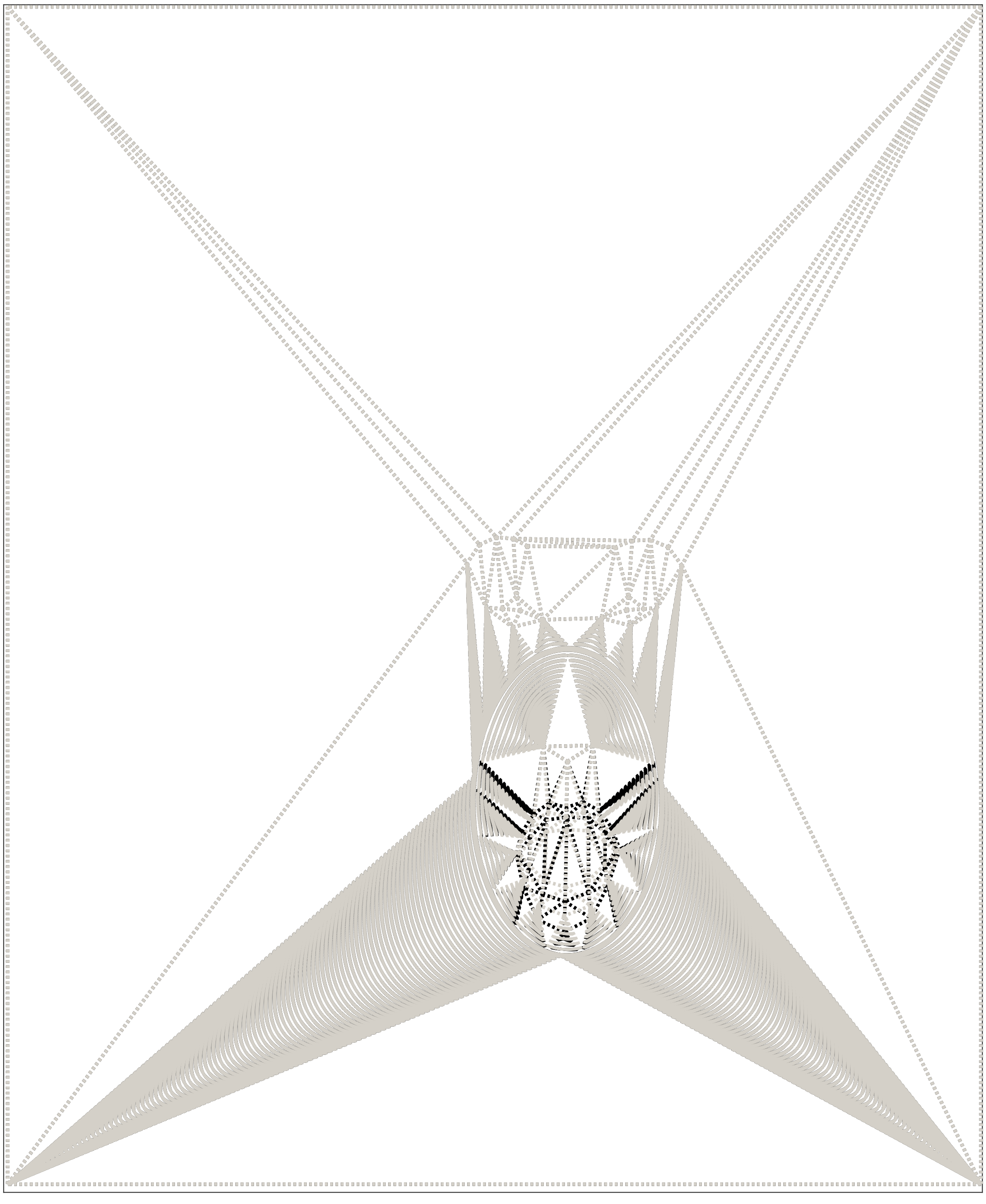


Figure 6.7: Triangulation of the control points of frame 28 using the rings; Gray: ***basline***, Black: ***exag***.

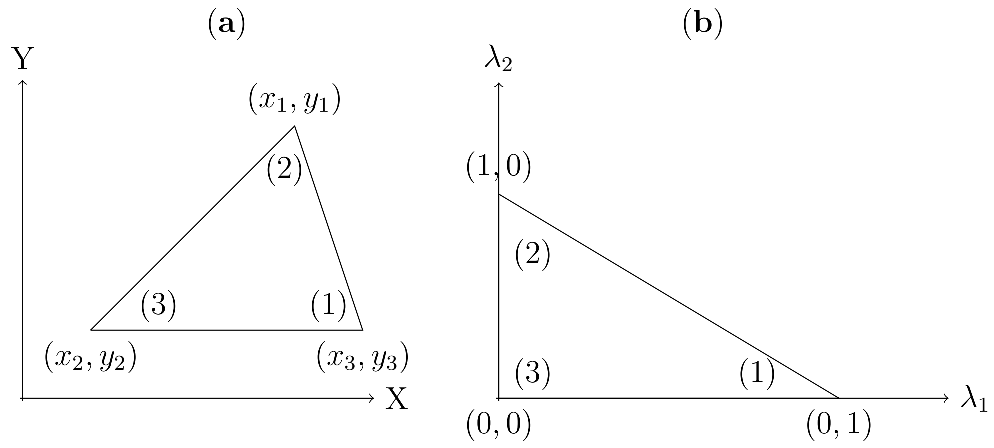


Figure 6.8: Transformation from (a) Cartesian to (b) Barycentric coordinate system, and vice versa. Adapted from [232].

Two sets of control points are selected to feed the warping process: the baseline set, $\boldsymbol{baseline}$, and the exaggerated set, $\boldsymbol{exag}$; both contain the image control points before and after the exaggeration,respectively, and can be expressed as:

$$\boldsymbol{baseline} = \begin{bmatrix}\boldsymbol{lip_k} & \boldsymbol{face_k} & \boldsymbol{ring} & \boldsymbol{frame\_corners}\end{bmatrix}^T \tag{6.10}$$

$$\boldsymbol{exag} = \begin{bmatrix}\boldsymbol{newlip_k} & \boldsymbol{face_k} & \boldsymbol{ring} & \boldsymbol{frame\_corners}\end{bmatrix}^T \tag{6.11}$$

where $\boldsymbol{lip_k}$, $\boldsymbol{face_k}$, and $\boldsymbol{newlip_k}$ are defined by Equations 6.1, 6.2 and 6.9, respectively. $\boldsymbol{ring}$ is a vector of points that form the perimeter of an ellipse delimiting the mouth region (Figure 6.5a and b). The frame's image corners were added to ensure that triangulation is applied to all image pixels. The convex hull of a control point set was partitioned into triangles using the Delaunay Triangulation (DT) method. The fundamental property that guides DT is the empty circle rule: the circumcircle of each triangle in the triangulation should be empty with no points inside [322]. Figure 6.5 illustrates the triangulation of the control point sets (image corners were excluded from the control point sets in this image to provide a close up look at the face mesh). Figures 6.6 and 6.7 show the impact of including $\boldsymbol{ring}$ in the control points. $\boldsymbol{ring}$ helped to create a region of interest where the exaggeration is applied and restricted. It also controlled the undesired propagation of the exaggeration effect as a result of the triangulation of the face area and the image background.

Inverse warping was then applied. Points inside a source triangle $\Delta v_1 v_2 v_3 \in \boldsymbol{exag}$ where $v_1 = (x\_eg_1, y\_eg_1)$, $v_2 = (x\_eg_2, y\_eg_2)$ and $v_3 = (x\_eg_3, y\_eg_3)$ were mapped to their corresponding points in a destination triangle $\Delta v'_1 v'_2 v'_3 \in \boldsymbol{baseline}$ where $v'_1 = (x\_bl_1, y\_bl_1)$, $v'_2 = (x\_bl_2, y\_bl_2)$ and $v'_3 = (x\_bl_3, y\_bl_3)$; eg denotes exaggerated

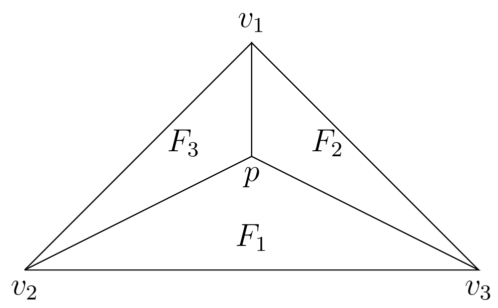


Figure 6.9: *Barycentric coordinates.*

coordinates and *bl* denotes baseline coordinates. This mapping function transforms source points from the *Cartesian coordinates system* to the *Barycentric coordinates* system, which is a universal coordinate system that describes points inside triangles using common features that can be preserved across triangles irrespective of their geometry (Figure 6.8). Using Barycentric coordinates means every point in a triangle is treated as a geometric centroid of three masses placed at the triangle vertices. These masses ($\lambda 1$, $\lambda 2$ and $\lambda 3$) are referred to as BC coordinates. BC coordinates of point $p = (x\_eg, y\_eg)$ in $\Delta v_1 v_2 v_3 \in \boldsymbol{exag}$ (Figure 6.9) can be defined as

$$\lambda_i = \frac{F_1}{F_1 + F_2 + F_3}, \quad i \in [1, 2, 3] \tag{6.12}$$

where F1, F2, and F3 are the areas of the sub triangles $\Delta p v_2 v_3$, $\Delta p v_3 v_1$ and $\Delta p v_1 v_2$ respectively. An important property of Barycentric coordinates is

$$\sum_{1}^{3} \lambda_i = 1, \quad \lambda_1, \lambda_2, \lambda_3 \geq 0 \tag{6.13}$$

Using Barycentric coordinates, $p = (x\_eg, y\_eg)$ can then be defined as

$$\begin{bmatrix} x\_eg \\ y\_eg \end{bmatrix} = \begin{bmatrix} \lambda_1 x\_eg_1 + \lambda_2 x\_eg_2 + \lambda_3 x\_eg_3 \\ \lambda_1 y\_eg_1 + \lambda_2 y\_eg_2 + \lambda_3 y\_eg_3 \end{bmatrix} \tag{6.14}$$

Using Equation 6.13, Equation 6.14 can be re-written as

$$\begin{bmatrix} x\_eg \\ y\_eg \end{bmatrix} = \begin{bmatrix} \lambda_1 x\_eg_1 + \lambda_2 x\_eg_2 + (1 - \lambda_1 - \lambda_2) x\_eg_3 \\ \lambda_1 y\_eg_1 + \lambda_2 y\_eg_2 + (1 - \lambda_1 - \lambda_2) y\_eg_3 \end{bmatrix} \tag{6.15}$$

and then to

$$\begin{bmatrix} \lambda_1 & \lambda_2 \end{bmatrix}^T = T^{-1}(x\_eg - x\_eg_3) \tag{6.16}$$

where

$$T = \begin{bmatrix} x\_eg_1 - x\_eg_3 & x\_eg_2 - x\_eg_3 \\ y\_eg_1 - y\_eg_3 & y\_eg_2 - y\_eg_3 \end{bmatrix} \tag{6.17}$$

Given this, λ_1, λ_2 and λ_3 can be expressed as

$$\lambda_1 = \frac{(y\_eg_2 - y\_eg_3)(x\_eg - x\_eg_3) + (x\_eg_3 - x\_eg_2)(y\_eg - y\_eg_3)}{det(T)} \tag{6.18}$$

$$\lambda_2 = \frac{(y\_eg_3 - y\_eg_1)(x\_eg - x\_eg_3) + (x\_eg_1 - x\_eg_3)(y\_eg - y\_eg_3)}{det(T)} \tag{6.19}$$

$$\lambda_3 = 1 - \lambda_1 - \lambda_2 \tag{6.20}$$

where $det(T)$ is the determinant of matrix T given by $((y\_eg_2\text{-}y\_eg_3)(x\_eg_1\text{-}x\_eg_3)\text{-}(x\_eg_2\text{-}x\_eg_3)(y\_eg_1\text{-}y\_eg_3))$. λ_1, λ_2 and λ_3, the Barycentric coordinate of point p, can then be used to infer the corresponding $p' = (x\_bl, y\_bl)$ in triangle $\Delta v_1' v_2' v_3 \in \boldsymbol{baseline}$ as

$$p' = \lambda_1 v_1' + \lambda_2 v_2' + \lambda_3 v_3' \tag{6.21}$$

After mapping all points bounded by a source triangle $\in$ $\boldsymbol{exag}$ to their corresponding points in a destination triangle $\in \boldsymbol{baseline}$, a bilinear interpolation that re-samples and interpolates pixels in the destination triangle is applied. For a pixel $P = f(x\_bl, y\_bl)$, $Q_{aa} = f(x\_bl_a, y\_bl_a)$, $Q_{ab} = f(x\_bl_a, y\_bl_b)$, $Q_{ba} = f(x\_bl_b, \_bl_a)$ and $Q_{bb} = f(x\_bl_b, y\_bl_b)$, which are the nearest four pixels to P, are used to interpolate the value of P (Figure 6.10). To achieve this, R_1 (the weighted average of Q_{aa} and Q_{ba}) and R_2 (the weighted average of Q_{ab} and Q_{bb}), are calculated as follows:

$$R1 = \frac{x\_bl_b - x\_bl}{x\_bl_b - x\_bl_a} Q_{aa} + \frac{x\_bl - x\_bl_a}{x\_bl_b - x\_bl_b} Q_{ba} \tag{6.22}$$

$$R2 = \frac{x\_bl_b - x\_bl}{x\_bl_b - x\_bl_a} Q_{ab} + \frac{x\_bl - x\_bl_a}{x\_bl_b - x\_bl_a} Q_{bb} \tag{6.23}$$

P can be then derived using the following equation

$$P = \frac{y\_bl_b - y\_bl}{y\_bl_b - y\_bl_a} R_1 + \frac{y\_bl - y\_bl_a}{y\_bl_b - y\_bl_a} R_2 \tag{6.24}$$

Note that for an RGB image, this process is repeated for every colour channel. Algorithm 1 outlines the major steps in the image warping process. Figure 6.11 shows exaggerated frames using $para = 1.5$ and 2 against the baseline frame. Figure 6.13 shows the visual exaggeration effects on viseme (visual phonemes) classes extracted from videos of one speaker from the GRID dataset [14]; the first column represents

Algorithm 1 Piecewise Inverse Image Warping

function DELAUNAY_TRIANGULATION (V)

Input: *V: a set vertices*

Output: *Triangulation: the set of triangles that make up the triangulation; each triangle is represented by its vertices indices V*

  Select $v_a \in V$ where $v_a(y)$ is the maximum ▷ The rightmost point in Pt
  Select $v_b, v_c \notin Pt$ such that the triangle $v_a v_b v_c$ contains V
  $Triangulation = v_a v_b v_c$ ▷ Initialise the triangulation
  for each v in V **do**
    for each T in $Triangulation$ **do**
      if v is inside T's circumcircle **then**
        $buffer = add\_new\_edge(buffer, T_{Edges})$ ▷ Add T's edges to buffer
        Delete (T)
    $buffer \leftarrow Unqiue(buffer)$ ▷ Remove duplicated edges
    for each $Edge$ in $buffer$ **do**
      $add\_new\_triangle(Triangulation, Edge, v)$
  for each T in $Triangulation$ **do**
    if v_{-1}, v_{-2} are vertices of T **then**
      Delete (T)
  return $Triangulation$

function IMAGE_WARP(I,xy,uv)

Input: *I: input frame image; xy: baseline control points; uv: exaggerated control points*

Output: *J: Exaggerated frame image is generated*

  $TRI = Delaunay\_triangulation(xy)$
  $J = I$;
  for each T in TRI **do**
    $V_{xy} = xy(T)$ ▷ Get baseline triangle vertices
    $V_{uv} = uv(T)$ ▷ Get exaggerated triangle vertices
    initialise $list$
    for each p inside T **do**
      $\lambda = Barycentric\_coordinate(p, V_{uv})$
      $p' = \lambda.(V_{xy})$
      $list \leftarrow p'$
    $bilinear\_interpolation(J, list)$
  return J

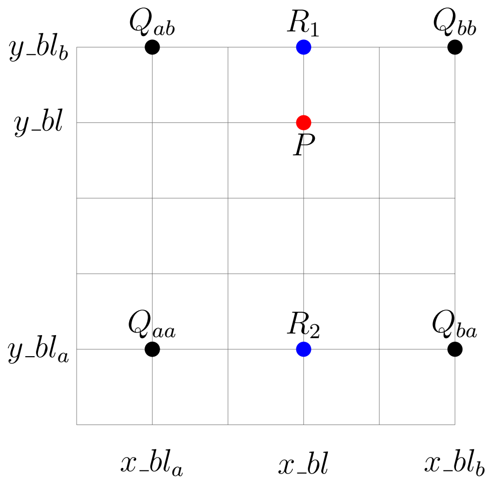


Figure 6.10: Bilinear interpolation; four points are used to compute the interpolation value P at (x_bl, y_bl).

the original viseme shape, while the second column represents the viseme after exaggeration by $para = 2$. In order to create a lipstick effect on the exaggerated video, a similar process as detailed in Section 5.3 was applied to the exaggerated videos (Figure 6.12).

6.2.3 Limitations

The teeth area is stretched as a result of the exaggeration effect (Figure 6.13). This stretching effect could be considered desirable, as it creates the illusion of showing more teeth and gums when the mouth aperture increases vertically or horizontally. However, when a talker has prominent teeth, this stretching may create an uncanny valley effect [229]. To exclude the teeth area from the exaggeration effect, teeth detection is required. To achieve this, K-means clustering [199] was used to segment the inner mouth area into colour clusters to locate the teeth area. K-means clustering partitions n data elements pi (pixels intensity colours) into k clusters by finding the positions of clusters that minimise the Euclidean distance between the data points within-cluster as follows:

$$\arg\min_{c} \sum_{n=1}^{k} \sum_{x \in ci} d(pi, \mu_i) = \arg\min_{c} \sum_{n=1}^{k} \sum_{pi \in ci} ||pi - \mu_i||_2^2 \tag{6.25}$$

where c_i is a set of points in cluster$_i$, μ_i is the mean of the points in c_i, and $||pi-\mu_i||_2^2$ is the square of the Euclidean distance of data in c_i. The K-means clustering algorithm begins by assigning μ_i to random values. The algorithm then runs iteratively over

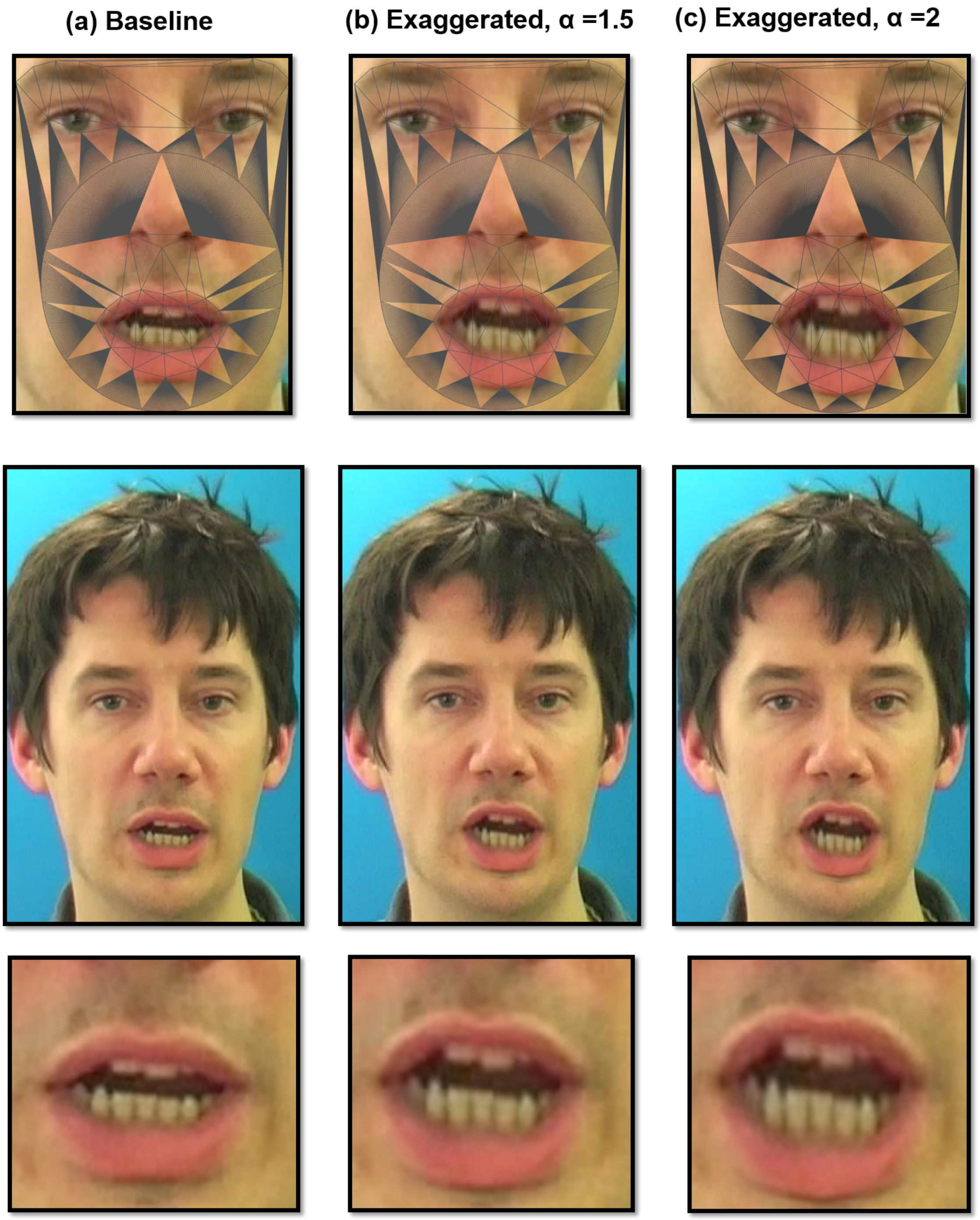


Figure 6.11: Frame warping after estimating exaggerated mouth shapes: (a) the original frame (frame 28); (b) and (c) frames under two levels of exaggeration effect ($para = 1.5$ and 2, respectively)

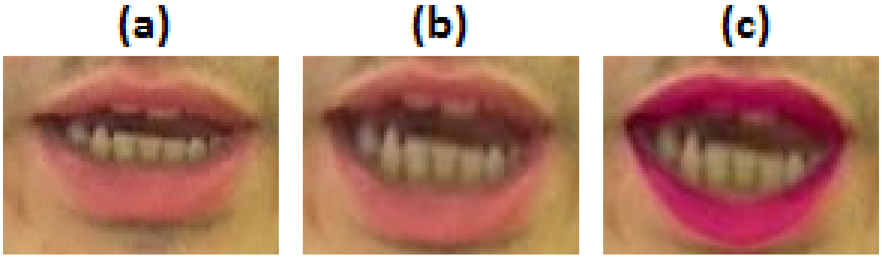


Figure 6.12: (a) A frame in the V stimuli; the corresponding frames in (b) E_2 and (c) E_3; $para = 2$ was used in the exaggeration effect applied to E and $E_{lipstick}$

two steps: the assignment step and the update step. In the assignment step, data point x is assigned to a c_i with the nearest mean. In the update step, μ_i is updated to match the mean of c_i as follows:

$$\mu_i = \frac{\sum_{j \in c_i} pi_j}{|c_i|}, \forall i \tag{6.26}$$

where $|c_i|$ is the size of c_i. The iteration of these steps is terminated when the convergence criteria is satisfied: no change is recorded for the assignment step [197].

To enable the Euclidean distance between pixels to be quantified, the frame image was converted from RGB colour space to CIE [2] $l\alpha\beta$ colour space prior to applying K-mean's clustering. CIE space describes colours that can be perceived by the human eye as tristimuli XYZ values, where Z defines luminance and XY defines the chromaticities of Z. The CIE $l\alpha\beta$ color space is derived from the XYZ color space, however, CIE $l\alpha\beta$ is more perceptually uniform to human visual perception (i.e., the Euclidean distance between two colours is strongly correlated with the perceived color differences) [310]. The CIE $l\alpha\beta$ colour space dimensions are 'l' for luminance, 'α' for colour value in the red-green axis, and 'β' for colour value in the blue-yellow axis.

Converting a frame image from RGB colour space to $l\alpha\beta$ colour space involves three steps [257]. First, a conversion from RGB in nominal range [0,1] to XYZ tristimuli values is applied as follows:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} .5141 & .3239 & .1604 \\ .2651 & .6702 & .0641 \\ .0241 & .1228 & .8444 \end{bmatrix} * \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{6.27}$$

In the second step, the frame image is converted from XYZ colour space to LMS colour space. LMS space represents a colour as the response of light by the three classes of human eye cones. The first class, L, responds to long light wavelength;

[2] Commission Internationale de l'Eclairage- International Commission on Illumination.

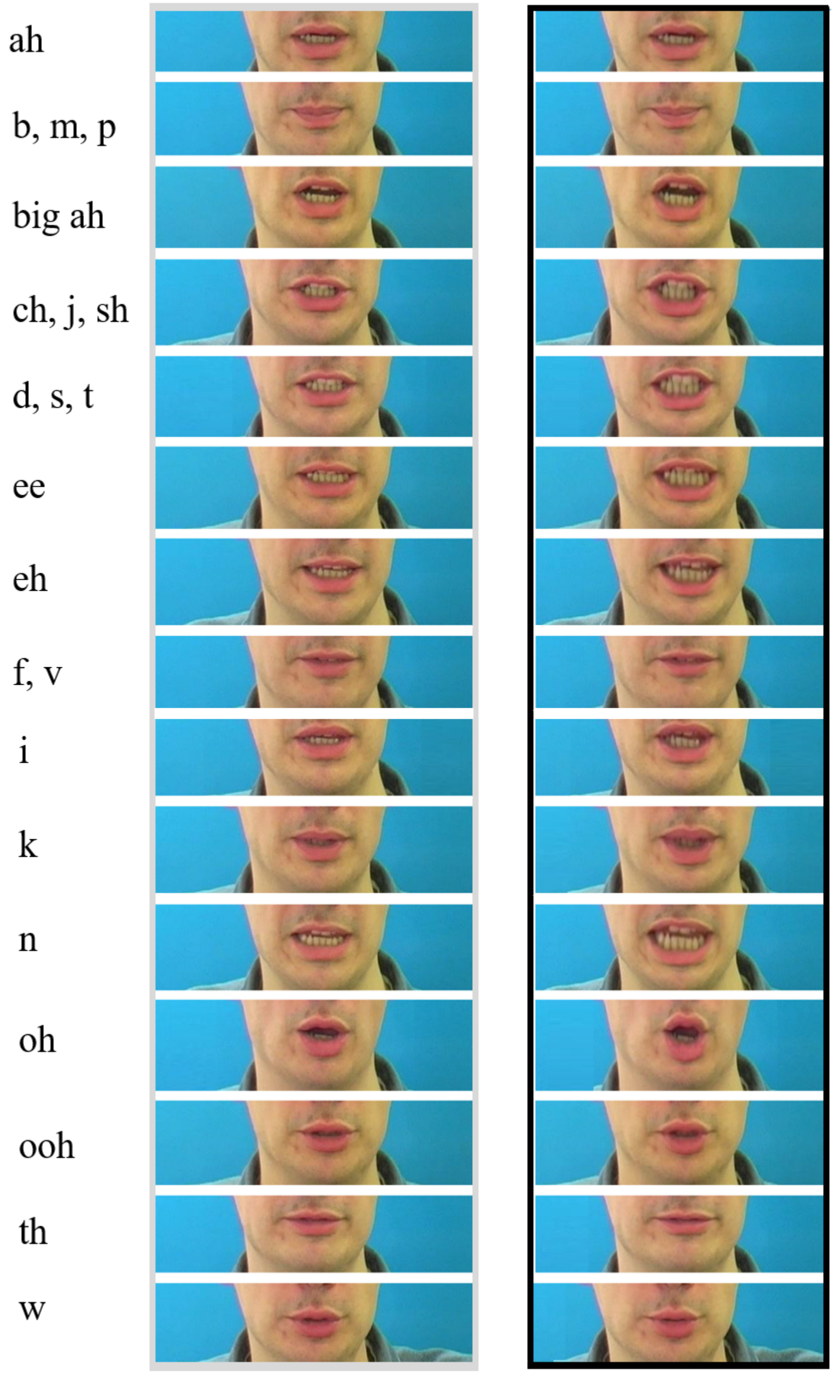


Figure 6.13: Viseme classes extracted for the Grid dataset [14]. The first column represents the original viseme mouth shape while the second column represents the viseme mouth shape after applying the exaggeration effect ($para = 2$). The British English Example Pronunciation dictionary was used for the phoneme notation.

the second class, M, responds to light of medium wavelength; and the third class, S, responds to light with short wavelength [142]. This conversion step is applied as follows:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.3897 & 0.6890 & -0.0787 \\ -0.2298 & 1.1834 & 0.0464 \\ 0.0000 & 0.0000 & 1.0000 \end{bmatrix} * \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{6.28}$$

The final step is to convert the frame image from LMS space to $l\alpha\beta$ colour space as follows:

$$\begin{bmatrix} l \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} * \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -2 \\ 1 & -1 & 0 \end{bmatrix} * \begin{bmatrix} L \\ M \\ S \end{bmatrix} \tag{6.29}$$

Colour data in $\alpha\beta$ space is then submitted to the clustering equation, Equation 6.25. For each frame, three clusters' RGB images were extracted: one containing the teeth, one containing the lips, and one containing the tongue and the gum region. In order to automate the process of selecting the right cluster that contains the teeth, all the resulted clusters' RGB images were converted to gray scale images. The cluster with a gray scale image that contains the brightest pixel intensity levels, i.e. the teeth region, was selected. Edge detection was then applied to that cluster to identify the teeth area contour in frame k, $\boldsymbol{teeth_k}$. The edge detection method used Sobel [287] operators to estimate the gradient components of the given image. The warping method is then applied after updating Equation 6.10 and 6.11 to include $\boldsymbol{teeth_k}$ so that the teeth area is excluded from the exaggeration as follows:

$$\boldsymbol{baseline} = \begin{bmatrix} \boldsymbol{teeth_k} & \boldsymbol{lip_k} & \boldsymbol{face_k} & \boldsymbol{ring} & \boldsymbol{frame\_corners} \end{bmatrix}^T \tag{6.30}$$

$$\boldsymbol{exag} = \begin{bmatrix} \boldsymbol{teeth_k} & \boldsymbol{newlip_k} & \boldsymbol{face_k} & \boldsymbol{ring} & \boldsymbol{frame\_corners} \end{bmatrix}^T \tag{6.31}$$

Figure 6.14 shows the detection results. Figure 6.15 shows the exaggeration result, with and without consideration of the teeth area during the warping process. Although the effect of 'prominent teeth' looks significantly improved at the single image level, a very noticeable jitter in the teeth area is observed in the video, even after smoothing the teeth tracking points. This suggests that, for better results, a video tracking method that takes into consideration mouth motion is required. As such jitter may contribute in creating an uncanny valley effect, the talker teeth in the exaggerated frames remained unadjusted.

The following section presents a study that tested the impact of using the exaggeration effect on the audiovisual stimuli used in auditory training. The

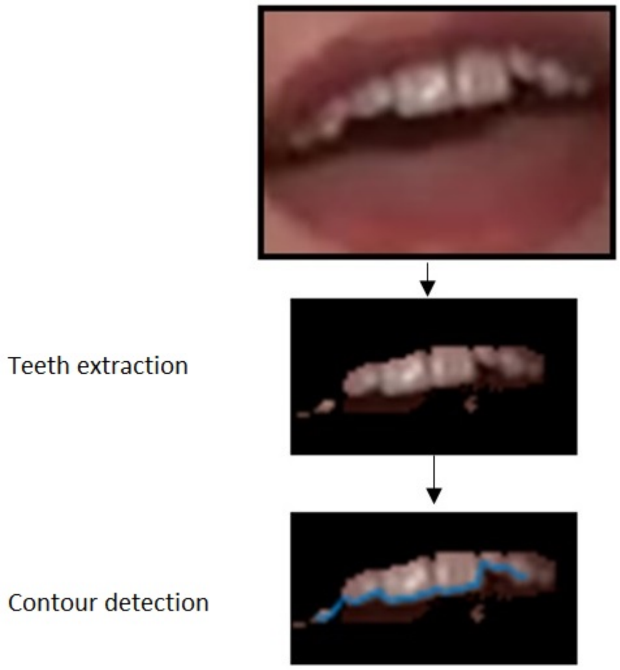


Figure 6.14: Detecting and extracting the teeth area.

audiovisual training framework, with adaptations, is used to test this impact of exaggeration on improving the intelligibility of CI simulated speech during and after the auditory training.

6.3 Evaluation study

6.3.1 The Audiovisual Training Framework

The audiovisual training framework (Section 4.4) was used to provide the evaluation for the effectiveness of the exaggeration effect. The framework follows the same methodology as the evaluation study (Section 4.4.1). This includes the use of the Grid corpus to provide the training stimuli, training methodology (an AO pre-test, 3 training sessions, an AO post-test) and baseline training modalities – A training that uses AO stimuli and V training that uses AV stimuli. The audiovisual training framework was adapted to introduce two more training modalities:

- E_2 training that uses AV_E stimuli – one level of exaggeration ($para = 2$) was applied on the AV stimuli to create AV_E stimuli, where subscript E denotes Exaggerated;

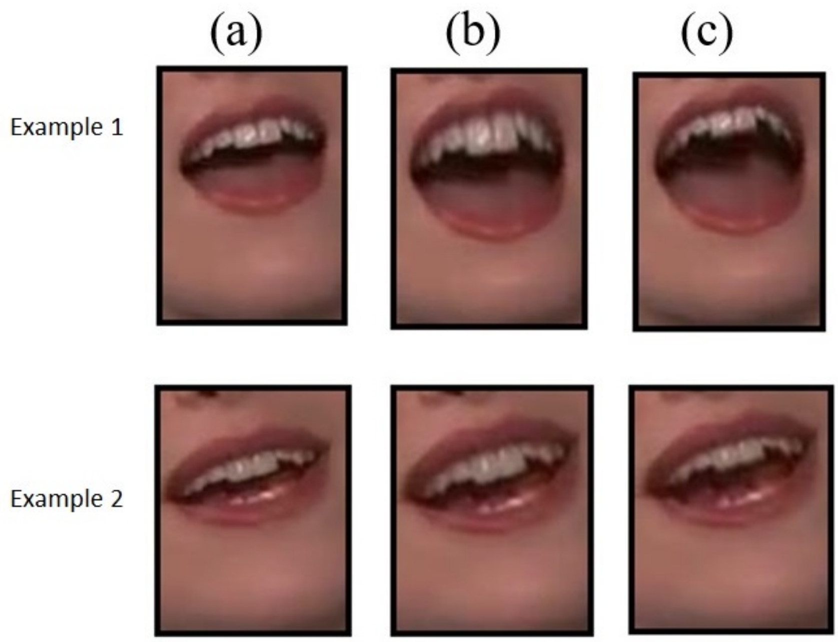


Figure 6.15: (a) Baseline frame (b) Exaggerated frame (c) Exaggerated frame with restriction of the teeth area.

- E_3 stimuli that uses AV_{LE} – one level of exaggeration ($para = 2$) was applied on AV_L stimuli (Section 5.3) to create AV_{LE} stimuli, where subscripts L and LE denote Lipstick and Exaggerated with lipstick applied, respectively.

Figure 6.16 and Table 6.1 summarise the training methodology and modalities used to evaluate the exaggeration effect. The recognition scores of subjects attending A and V training will be used to evaluate the effect of the exaggeration and the combined effect of the lipstick and exaggeration, i.e., the recognition scores of subjects attending E_2 and E_3 training.

Another modification to the audiovisual training framework is the process of assigning subjects to subgroups. As noted in Section 5.4, the random allocation of subjects created subgroups with variable baseline levels. To resolve this issue, the assignment of a subject S to a training subgroup was done automatically when the subject finished the pre-test so as to establish a similar baseline across all subgroups (Figure 6.16). This was done as follows: Assume that the subject's pre-test score is $S_{pre-test}$. The training software finds a subgroup X (X = A, V, E_2, or E_3) such that adding $S_{pre-test}$ to the set of X pre-test scores minimises, makes no change, or makes the minimum increase to the standard deviation between the means of all subgroups' pre-test scores. After the assignment, all subjects resumed training and testing in a similar way to the lipstick experiment (Section 5.4); individuals were trained in

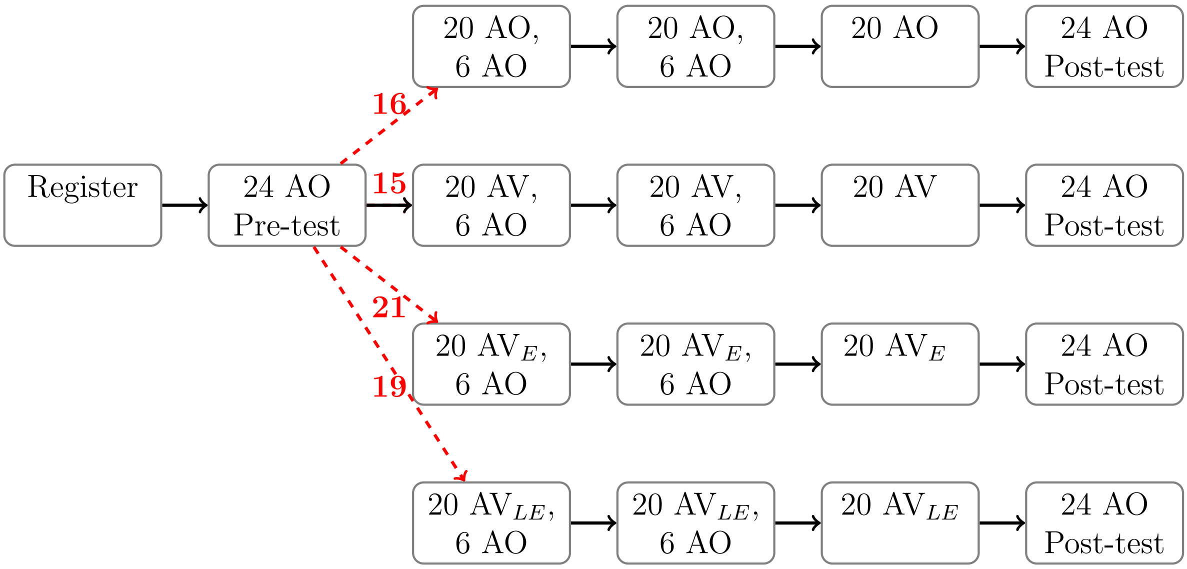


Figure 6.16: The training method consisted of an audio-only pre-test followed by three training sessions then an audio-only post-test session for determining the training gain. The numbers of subjects for each route through the training process are shown in red.

one of four alternative conditions: (1) audio only, (2) audiovisual, (3) exaggerated audiovisual, or (4) exaggerated audiovisual with simulated lipstick applied, and were then tested again using audio-only stimuli.

6.3.2 Subjects

The experiment was conducted at the female campus of King Saud University in Riyadh, Saudi Arabia. The subjects were 71 female non-native Saudi listeners (with minimum IELTS score = 5.5), each in the age range 18–40 years (mean = 24 years; standard deviation [SD] = 4.5 years). Ethics permission for this study was obtained by following the University of Sheffield Ethics Procedure. As a result of the automatic assignment of participants to groups, the participants were split into four groups: A (16 subjects), V (15 subjects), E_2 (21 subjects), and E_3 (19 subjects).

6.3.3 Results

Figure 6.17 summarises the main results of this experiment. Figure 6.17a shows a comparison of the four groups across all training sessions. Between groups, one-way ANOVA testing for the groups showed a significant difference between the V and E_2 groups during the second training session ($F(3, 67) = 3.38$, $p = 0.02$).

	Audio	Video
AO stimuli, A training	CI simulated Grid audio	-
AV stimuli, V training		Grid video
AV_E stimuli, E_2 training		Grid video with exaggeration effect
AV_{LE} stimuli, E_3 training		Grid video with exaggeration and lipstick effect

Table 6.1: Stimuli and training modalities. subscripts E, LE denote Exaggerated and Exaggerated with Lipstick applied, respectively.

No significant difference was found between other groups in all training sessions. Within groups, repeated-measure ANOVA showed a significant difference between sentence-recognition scores in the E_2 training sessions ($F(2, 40) = 9.987$, $p = 0.000$). A post-hoc pairwise comparison found a difference between sessions 1 and 3 ($p = 0.012$) and sessions 2 and 3 ($p = 0.000$). A significant difference was also found between the sentence-recognition scores in the E_3 training sessions ($F(2, 36) = 3.38$, $p = 0.02$); the post-hoc test demonstrated a difference between sessions 1 and 3 ($p = 0.038$). No significant difference was found between the sentence-recognition scores in all sessions of the A and V training. Subjects who underwent the E_2 and E_3 training described the modified form of speech as having incongruent audiovisual signals. More energy was observed in the visual signals than in the audio signals (i.e., the video cues were more salient than the audio cues). This situation made audiovisual signals unintelligible at the start of the training.

Figure 6.17b shows the mean sentence-recognition scores that the A, V, E_2, and E_3 subjects attained in their audio-only pre- and post-training tests as well as their mean training gains in auditory recognition. All subgroups were formed with comparable pre-test scores (as a result of the automatic assignment process). The V subjects achieved the highest post-test sentence recognition and training gains, although a one-way ANOVA test showed no significant difference between their sentence recognition scores in post-testing ($F(3, 67) = 1.6$, $p = .19$) among the A, V, E_2, and E_3 subjects. A Levene's test [182] indicated unequal variances in training-gain scores ($F = 2.6$, $p = .041$), while Welch ANOVA testing reported no significant difference in training gains among all groups ($F(3, 35) = 0.00$, $p = 1.00$). Figure 6.17c shows the scores for the stimuli used in the 6 AO of sessions 1 and 2 for all subgroups, which were used as Learning Milestones to track the audio-only skills for all subgroups throughout the training. As Figure 6.17c shows, no significant differences were found in the session 1 post-testing among all groups (Milestone 1; $F(3, 67) = 0.14$, $p = .93$), while a significant difference was found between the A and V groups in session 2 post-testing (Milestone 2; $F(3, 67) = 2.91$, $p = 0.04$, post-hoc

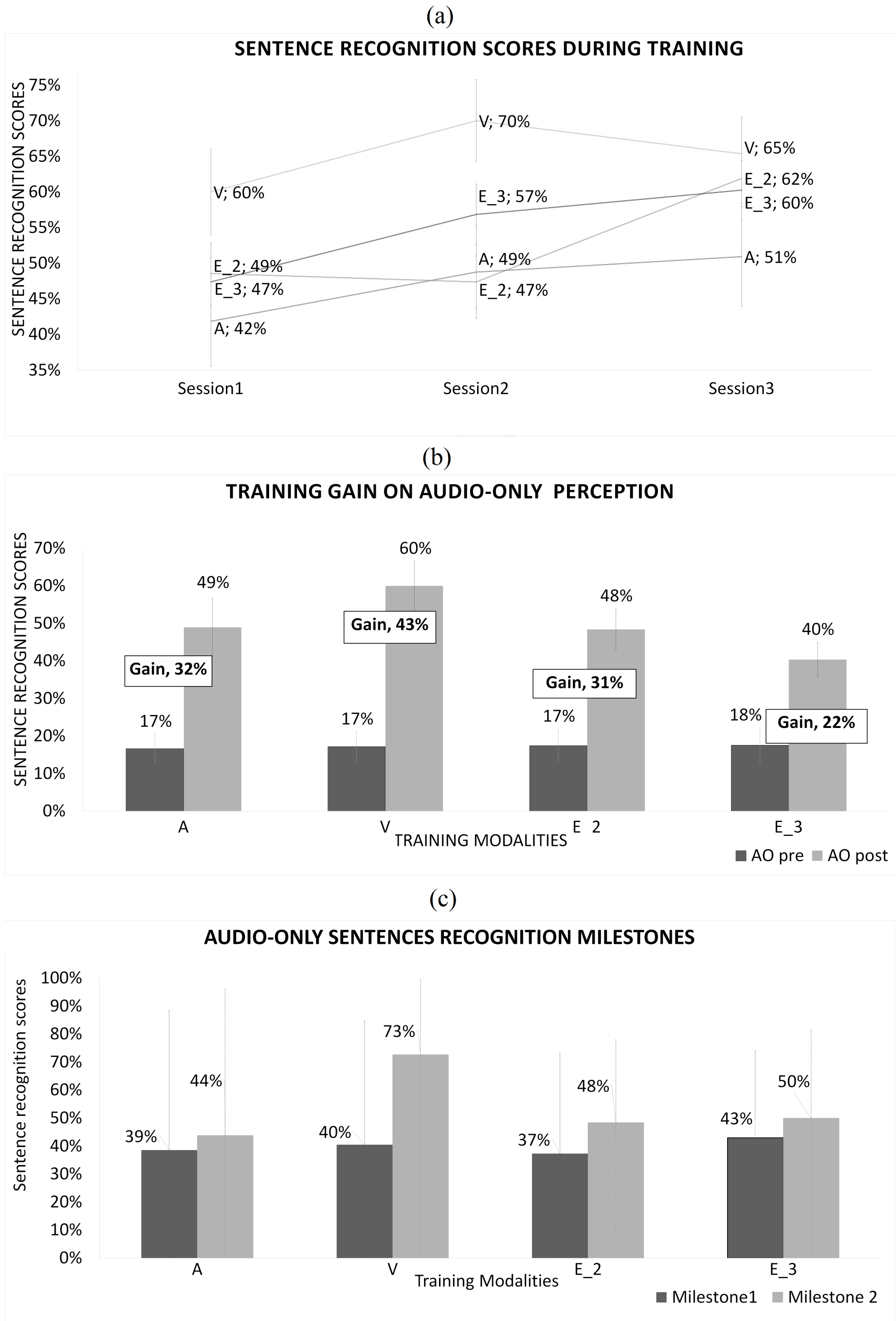


Figure 6.17: Results for the A, V, E_2, and E_3 subjects: (a) sentence recognition during training; (b) audio-only pre- and post-test mean identification scores and training gains (post-test results - pre-test results); (c) training impact on audio-only sentence recognition (learning milestones). Errors bars =/- standard error.

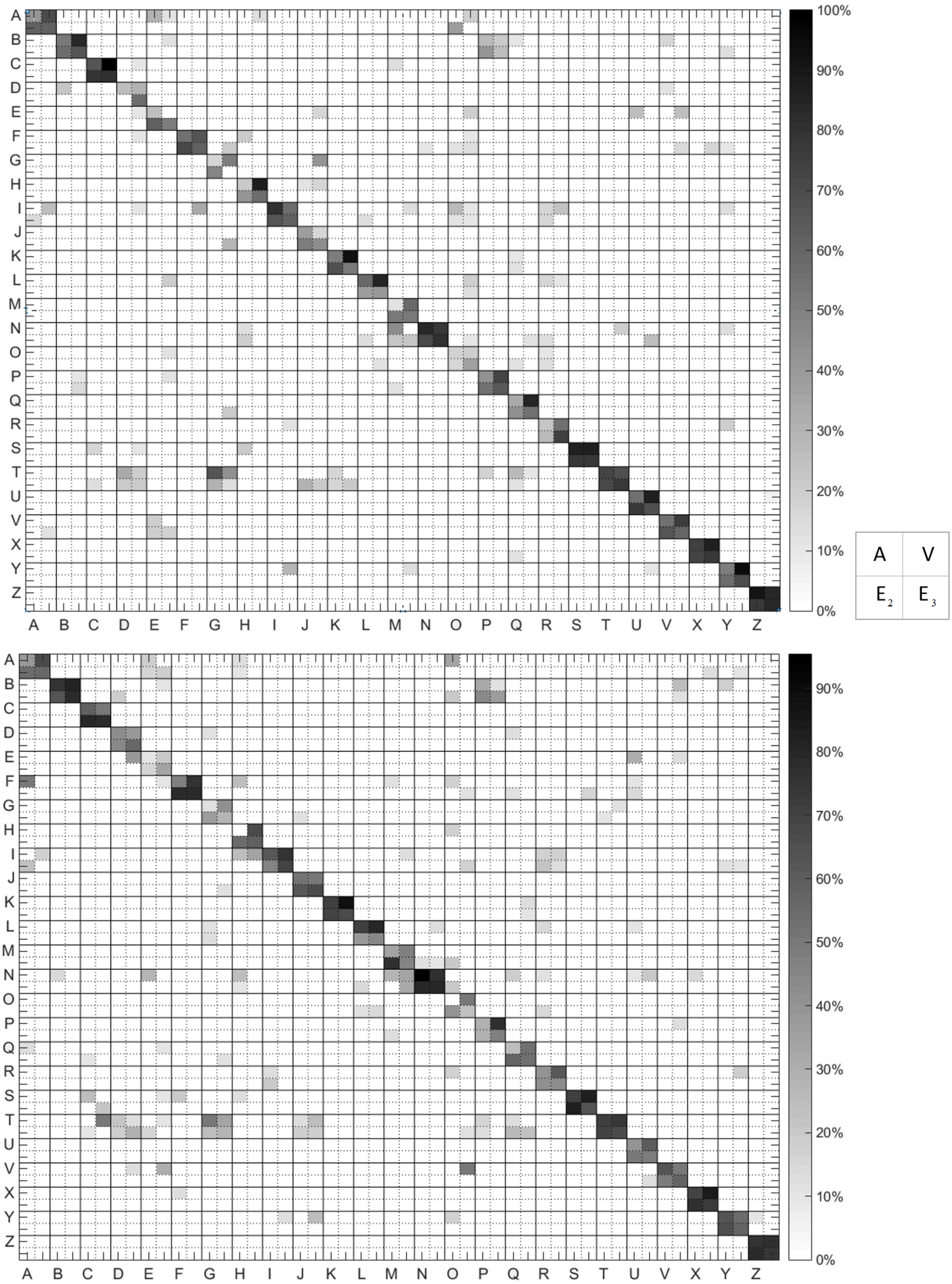


Figure 6.18: Confusion matrix of letter-recognition scores during audio-only training (top), post-testing (bottom) for A, V, E_2 and E_3 subjects. Each cell is divided into 4 sub-cells: A, V, E_3 and E_2 (in clockwise order). Colour shades represents the scale of recognition/confusion mean rate (the darker the shade, the higher the response to this cell).

Tukey HSD test $p < .05$). Only the V subjects showed significant improvements in their recognition scores between Milestones 1 and 2 ($p = .002$). This suggests that using unmodified visual signals in the training may help speed up the V subjects' learning curve for auditory-only skills.

Confusion matrices (Figure 6.18) were also produced in order to understand the possible sources of confusion the subjects experienced in letter keywords recognition during the training and audio-only post-testing. Letter recognition was found to be the most challenging task for all subjects due to the need to select from a larger set with high variance (25 letters), and letters having shorter duration in terms of phonemes, and thus less information, as opposed to colours (4) and digits (10). During the training, the introduction of either modified or original visual signals improved the subjects' recognition of letters that contained vowel sounds (as well as bilabial, labiodental, and velar consonant sounds) compared to the audio-only training regime. Modified visual signals (E_2 and E_3 stimuli), however, improved the vowel recognition by 10% and the recognition of letters containing alveolar consonants by 7%. The A subjects' high confusion rates were observed in the pairs (A /**e**ɪ/, E /**iː**/), and in (N /e**n**/,M /e**m**/) and (T /**t**iː/, D /**d**iː/- G /**dʒ**iː/). These high rates may have indicated low abilities in processing vowels, nasality, and voicing cues that were distorted by the vocoder. High confusion rates were also observed among the V subjects for pairs (G /dʒ**iː**/, J /dʒ**eɪ**/) and (I /**a**iː/, E /iː/) which are visually similar letters. The E_2 subjects' high confusion rates were observed for the pairs (A /**eɪ**/, O /**oʊ**/) and (P /**p**iː/,B /**b**iː/); this was likely the result of over-exaggerated mouth shapes.

After the training and during the audio-only post-testing, the V subjects outperformed all the other groups auditory recognition of vowel letters, since they recognised 54% of the vowels compared with the A, E_2, and E_3 subjects, which scored 30, 41, and 46%, respectively. No significant difference in consonant recognition was found among the groups. The A subjects showed confusion between the (G /**d**ʒiː/, T /**t**iː/), (V /viː/, O /oʊ/), (O /**oʊ**/, A /**eɪ**/) and (P /**p**iː/, B /**b**iː/) pairs; the V subjects with (C /siː/, T /tiː/), (T /**t**iː/, G /**d**ʒiː/) and (V /v**iː**/, E /**iː**/); and the E_2 and E_3 subjects with the (P /**p**iː/, B /**b**iː/) and (Q /kjuː/,T /tiː/) pairs.

6.3.4 Discussion

Chapter 5 examined the effectiveness of automatically enhancing the appearance of a talker's lips in order to maximise the benefit of visual speech for improving the intelligibility of CI simulated speech in audiovisual training. This chapter has gone further by investigating the impact of exaggerating a talker's mouth kinematics in audiovisual speech. Because visual signals are a correlate of audio signals in audiovisual speech, exaggerating the visual signal alone in audiovisual speech can create incongruent inputs for listeners. Given this situation, the study reported in this chapter investigated the subjects' ability to adapt to audiovisual mismatches after exaggerating visual speech. The study also investigated whether or not applying the exaggeration effect to audiovisual speech would improve the benefits of the visual signal. The main observation and discussion points are as follows.

Audiovisual Training Consistent with previous findings [15, 28, 156, 250], the introduction of unmodified audiovisual speech during auditory training was found to improve the training gains in the auditory and audiovisual perception of CI simulated speech. Visual speech facilitation for speech-in-noise intelligibility [296] played a key role in improving the non-native subjects' audiovisual recognition rates for the CI simulated speech during the training. Using visual speech in training improved the subjects' auditory adaptation processes to spectrally distorted speech; the subjects were found to have significantly improved between learning milestones. This situation could reflect the impact of effective rapid perceptual learning.

Audiovisual Conflict After-effect After exposure to the audiovisual speech with the exaggeration effect (E_2 and E_3), evidence was found of an audiovisual conflict after-effect. The subjects were sensitive to the conflict between the articulation energy and the vocal effort in the modified videos during the early training stages. They also underwent a recalibration process during audiovisual speech integration in order to overcome this conflict. This situation was supported by the adaptation profile of the modified audiovisual speech groups (Figure 6.17a), which showed a dramatic increase in the audiovisual recognition rate during session 3. The increase reached a comparable level to that of the group that received congruent audiovisual speech signals (the V group), which indicates that the conflict impact became negligible to the E_2 and E_3 subjects after exposure. There is a difference, however, observed in the pace of the adaptation process between the E_2 and E_3 subgroups; the E_3 subgroup seemed to adapt faster as reflected by the increase in the recognition scores between sessions 1 and 2 in Figure 6.17a. This suggests that the lipstick filter may have an impact on accelerating the adaptation process in the E_3 subjects.

Impact of the Exaggeration on Audiovisual and Auditory Recognition The exaggeration of the visual speech signal also improved the audiovisual recognition of vowels and alveolar consonants, which are included in 44% of the Grid letters. For the remainder of the Grid letters, the exaggeration of the visual signal showed a comparable benefit to the unmodified visual speech. However, it did not improve the subsequent auditory recognition. Those subjects who were trained with exaggerated speech attained training gains in auditory recognition that were comparable to the gains of those who had been trained with auditory-only speech. This situation indicates that the subjects did not make use of the modified visual signals to facilitate their auditory adaptation to the spectrally distorted speech. A hypothesis is that the recalibration process the subjects underwent in order to adapt to the audiovisual conflict during the training was responsible for this. The recalibration process may have introduced additional cognitive load to the subjects, which in turn slowed down their auditory perceptual learning. It is thus difficult to judge whether or not modifying the audiovisual speech by exaggerating the visual signal can maximise the training gains in auditory recognition, since the subjects needed to undergo a recalibration process in order to adapt to the modified signals before they commenced the training.

Improving the Exaggeration Effect Hyper-articulated speech is a more sophisticated phenomena than just a process of exaggeration. It is govern by rules, theory and frameworks that describe and regulate this effect. One important theory that describes the energy behaviour in hyper-articulated speech is the Hyper- and Hypo- articulation (H&H) theory, which suggests that hyper-articulation energy is not constant, but variable across time. In the presented exaggeration method, visual signals were amplified by a constant value for all speech segments, which contradicts the H&H theory. This suggest that, in order to accurately simulate the effect of exaggeration, understanding the mouth behaviour in real hyper-articulated speech is essential. There are many examples of hyper-articulated speech that can be addressed, including clear speech, infant-directed speech, non-native directed speech, and speech induced noise or Lombard speech [192]. Lombard speech is a convenient example of hyper-articulated speech, and can be chosen as a subject of an analysis study that characterises visual hyper-articulated speech. This is because Lombard speech can easily be induced in a controllable manner compared with other forms of hyper-articulated speech [283, 284]. One of the main obstacles in conducting such an analysis study is the lack of Lombard speech datasets. In Chapter 7, a new Lombard audiovisual dataset is presented, along with an analysis study that examines visual Lombard speech from different perspectives.

6.4 Summary

This chapter presented the implementation and evaluation of a kinematics based visual speech enhancement approach (the automatic exaggeration effect). The results of the evaluation study suggest that the subjects who attained the ability to adapt to the mismatch between visual and audio signals did so as an after-effect of exposure to the exaggeration of the visual signal of audiovisual speech. As audiovisual conflict became negligible to subjects' after exposure, the results suggest some potential in applying enhancement effects on the visual signal alone in audiovisual speech, even if such an enhancement may create incongruent audiovisual inputs, as this effect has improved the subjects' audiovisual recognition of certain phoneme classes.

The exaggeration effect, however, did not produce similar improvements in the subjects' subsequent auditory recognition. Their adaptation to the audiovisual conflict during the training may have played a role in impeding their use of the visual signal in improving their subsequent auditory-only skills. The next chapter will examine visual speech enhancement in the real-life phenomenon of hyper-articulated speech, that is Lombard speech.

Chapter 7

Visual Lombard speech analysis

7.1 Introduction

In Chapter 6, a kinematics based enhancement approach that automatically exaggerates a talker's speaking style was presented. This technique simply amplified mouth movement. In contrast, this chapter will investigate Lombard speech [192], a real-life example of exaggerated or hyper-articulated speech. Lombard speech is accompanied by a set of acoustic, phonetic and articulatory adaptations that are associated with increased intelligibility [13, 63, 92, 161] (See section 2.5.4). For this study, visual Lombard speech has been chosen as a case study for understanding visual speech enhancements in real life since it can be easily induced and controlled due to its reflexive nature. Global adaptations of visual Lombard speech have received considerable attention in the literature (see Section 2.5.4). However, very few studies have focused on adaptations at the phoneme level [100, 102]. This chapter therefore presents a study of adaptations of visual Lombard speech at the utterance level and phoneme level within different contexts. The results of this study provide an increased understanding of visual speech enhancements associated with hyper-articulation and could be used to model the kinematics of the articulatory features observed in visual Lombard speech. The results could therefore improve the exaggeration effect described in Chapter 6.

To undertake this analysis, Lombard speech data that is recorded in a controlled environment are needed. One thing that holds back research on Lombard speech is the lack of suitable datasets that fulfill this requirement (see section 2.5.4). Therefore, an audiovisual dataset of Lombard speech based on the widely used audiovisual Grid corpus [54] was recorded under a high SNR level whereas listeners were exposed to low SNR via headphones. This dataset offers a plain (non-Lombard) reference for each recorded Lombard sentence. It also features two synchronised views of the talker – a front view and a side view – which enables the visual Lombard speech to be characterised from different angles. The video recordings were made using head-mounted cameras to stabilise the talker's head pose throughout the recording and therefore allow precise comparison of the Lombard and plain utterances.

This chapter is organised as follows. The audiovisual Lombard dataset is presented in Section 7.2. The audiovisual equipment used in the dataset recording is presented in Section 7.2.3 The dataset collection procedure is presented in Section 7.2.4. The post-processing of the dataset into a format suitable for analysis is presented in Section 7.2.5. Finally, an analysis of the visual Lombard speech is presented in Section 7.3.

7.2 A Corpus of Audiovisual Lombard Speech with Front and Profile Views

7.2.1 A Population of Talkers

The talkers that participated in the experiment were 55 native speakers of British English (both male and female) who were all staff or students at the University of Sheffield, each in the age range 18 – 30 years. The hearing of the talkers was screened using an on-line pure tone audiometric test [249]. Participants were paid for their contributions. Ethics permission for this study was obtained by following the University of Sheffield Ethics Procedure.

7.2.2 Sentence and Masker Design

The recording sentences conformed to the Grid corpus [54] syntax (see Section 4.4.1). A sentence in the Grid syntax, such as 'bin blue at A 2 please', consists of a six-word sequence with the following structure: <command: 4> <color: 4> <preposition: 4> <letter: 25> <digit: 10> <adverb: 4>. The number of choices for each keyword is indicated by the number in the angled brackets (Table 4.2). Three of these words – colour, letter and digit – were chosen as 'keywords' and the remaining were used as 'fillers'.

The original Grid [54] corpus was collected from 34 talkers reading 34,000 sentences selected from 64,000 possible combinations of the Grid word sequences. For the new Lombard Grid corpus, 55 talkers uttered sets of sentences selected from the pool of the remaining 30,000 Grid word sequence combinations; i.e., those that were not used in the original Grid corpus. The sets used in this thesis are listed below.

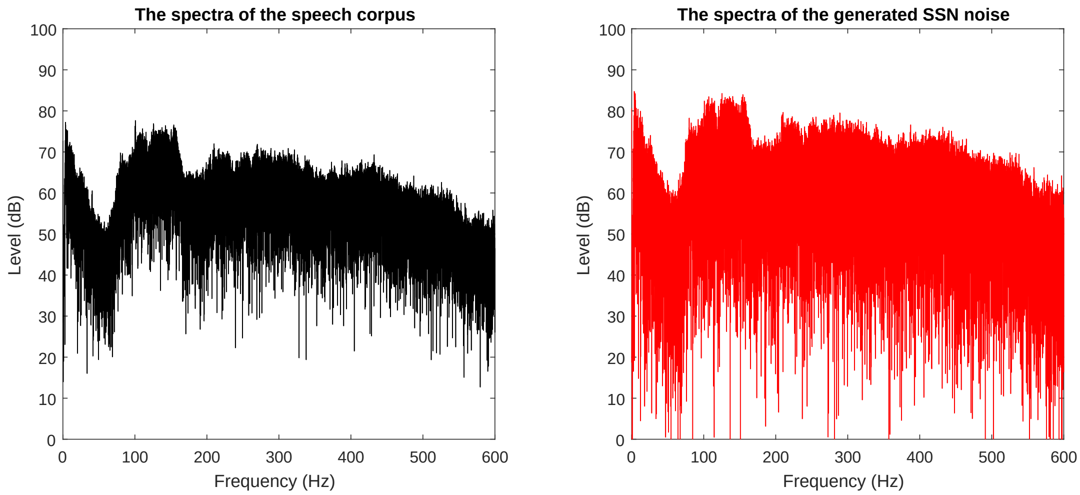


Figure 7.1: The spectra of (a) the speech corpus and (b) the generated SSN noise

- *55 actual sentence sets*: These sets were called 'actual' sentence sets since the talkers' utterances made from these sentences were meant to appear in the dataset. Each talker was assigned to a unique actual sentence set of size 50 sentences. Thus, there was a total of 55 $\times$ 50 sentence sets.

- *One 'warm up' sentence set*: This 50-sentence set was read by all the talkers, and the utterances made from this set were discarded in the final production stage. The sentences in this set were used to attune the talker's articulation during the transition from one experimental condition to another (e.g. from Lombard speech to plain speech).

An actual set featured a uniform representation of Grid keywords as follows:

- Twelve to fourteen instances of each colour. The distribution of colour instances followed two patterns: Two colours appeared 12 times and the other two colours appeared 13 times, or three colours appeared 12 times and the fourth colour appeared 14 times.

- Two instances of each letter;

- Five instances of each digit;

- A good coverage of the Grid filler words.

Speech-shaped noise (SSN) was used as a noise masker. The talkers were exposed to the SSN via a pair of headphones. This method was chosen because it provided

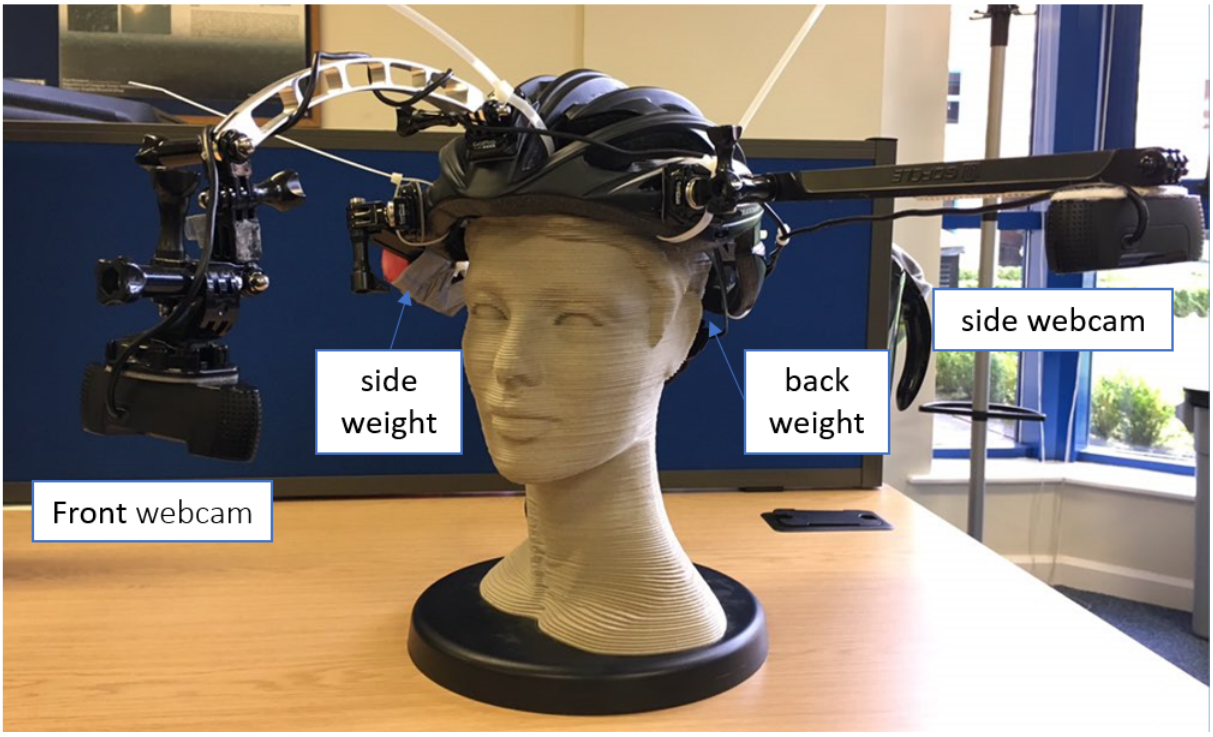


Figure 7.2: The recording helmet.

sufficient energetic masking of noise to induce Lombard speech [193, 194]. The SSN was created by filtering white noise to match the long-term spectrum of a speech corpus that included 1,000 Grid corpus [54] sentences of a selected talker (Grid corpus ID = 1). Linear predictive coding [241] was used as a filter which provided the spectral envelope of that speech corpus. Figure 7.1 illustrates the spectra of both the SSN and the speech corpus used to generate this noise.

In previous Lombard-related studies, maskers used to induce Lombard speech were presented at various levels, such as 80 dB SPL [164, 284, 301], 85 dB SPL [92, 152], and 89-96 dB SPL [194]. For this study, 80 dB SPL[1] was chosen as the masker noise level to minimise the impact of vocal and auditory fatigue on the talkers who would be exposed to high sound pressure levels.

7.2.3 Recording Equipment

Audio

The recordings were made in a single-walled acoustically isolated booth (Industrial Acoustics Company – IAC). The speech material was collected at a sampling rate of

[1]80 dB SPL is within the acceptable ranges for the daily exposure according to the Health and Safety Executive Organisation. [2]

48,000 Hz and a sample format of 24 bits using a C414 B-XLS AKG microphone placed 30 cm in front of the talker and digitised using the MOTU 8-pre 16×12 Audio Interface. The talkers wore Sennheiser HD 380 pro headphones. The SSN masker was mixed with the audio signal of their speech to give self-monitoring feedback at a level that compensated for headphone attenuation. The reason for the self-monitoring feedback was to reduce the potential increased speech modification resulting from wearing closed headphones [102]. The level of playback of the talker's speech was carefully adjusted so that their perceptions of talking with and without the headphones were comparable. The level of the masker pressure was calibrated against an SPL meter (see Section A.1 for further details). The collection of the talker's speech was controlled by a computer (Computer A – a Mac Mini; processor: 2.6 GHz Intel Core i5; memory: 8 GB 1600 MHz DDR3) connected to the audio interface using Audacity software [308]. The masker presentation was controlled by another computer (Computer B – a MacBook Pro; processor: 2.9 GHz Intel Core i5; memory: 8 GB 1867 MHz DDR3; USB 3.1) which was also connected to the audio interface.

Video

In addition to the audio recordings, simultaneous audiovisual recordings were made using a bespoke helmet rig system worn by the talkers (Figure 7.2). The system included two webcams that captured front and profile views of the talkers. The system consisted of a lightweight bicycle helmet with two Logitech HD Pro USB Webcam C920s[3] connected to a GoPole[4] Arm Helmet Extension (8 inches) fitted to the helmet using 3M adhesive mounting tape. The first armature was attached to the front of the helmet and was connected to the front webcam using a GoPro[5] pivot arm. The second armature was fitted to left side of the helmet and was connected to the side webcam using double-sided adhesive tape. A dumbbell weight (0.5 g) was attached at the rear of the helmet to counterbalance the weight of the front camera. Another weight (0.5 g) was connected to an arm attached the right side of the helmet to counterbalance the weight of the side camera. The talkers wore a soft hat to cushion the helmet. This also helped to fix the helmet on the talker's head. After the helmet was placed on the talker's head, a pair of headphones were fitted behind the talker's head and then attached to the helmet using two self-locking nylon cable ties.

[3]Cameras' mounts were removed to reduce weight.
[4]https://www.gopole.com/
[5]https://gopro.com/

	Block 1					Break	Block 2				
Talker A	P	P	L	L	P		L	P	P	L	L
Talker B	L	L	P	P	L		P	L	L	P	P

Table 7.1: Recording schedules: P is a plain session, and L is a Lombard session. For each session, a talker reads a prompt list of 5 warm-up sentences followed by 10 actual sentences.

The audiovisual recordings from the webcams were collected using two computers; the webcams were connected to the machines via USB 2.0 extension cables. The audiovisual stream from the front webcam was collected using the Photo Boot app running on Computer B at 480p (720 x 480) and in full frame at a variable frame rate fluctuating around 24 frames per second (mean FPS = 23.93; mean bitrate = 2817.82 kb/s)[6]. The app encoded the video stream using the built-in H.264 encoder and the audio stream using the AAC encoder at a sampling rate of 44,100 Hz. The video stream from the side webcam was collected using Logitech software installed on Computer C (HP Envy; processor: Intel Core™ i7-4702MQ; memory: 16 GB; USB 3.0) at 480p (864 x 480) and in full frame at 30 FPS. The Logitech software encoded the video stream using the WMV encoder and the audio stream using wmav2 at a sampling rate of 48,000 Hz. Four light sources were placed in different locations to produce uniform illumination across the talker's face, and a plain white background was placed behind and at the right side of the talker's seat. Figure 7.3 shows example frames from recorded videos collected from the front and side cameras.

Prior to collecting the dataset, a pilot study was conducted to examine variables that could regulate the effect of the Lombard speech (see Appendix A.3 for more details). In this study, the impact of the recording duration, the number of masker presentation levels and the talker's task during the recording were analysed. The results of this pilot study informed the design of the collection procedures presented in the following section.

7.2.4 Collection

In this dataset, each talker produced 150 sentences by reading 10 prompt lists. The prompt lists were generated as shown in Figure 7.4: the actual sentence set (Section 7.2.2) given to each talker was shuffled and broken down into five prompt lists for the plain recording and then reshuffled and broken down into five prompt lists for

[6]Videos of the front camera were in variable frame rate as a result from the preemptive multitasking nature of the Macbook machine.

Figure 7.3: Selected samples from the dataset. Top to bottom talker ID: 55, 44, 46 and 32, respectively.

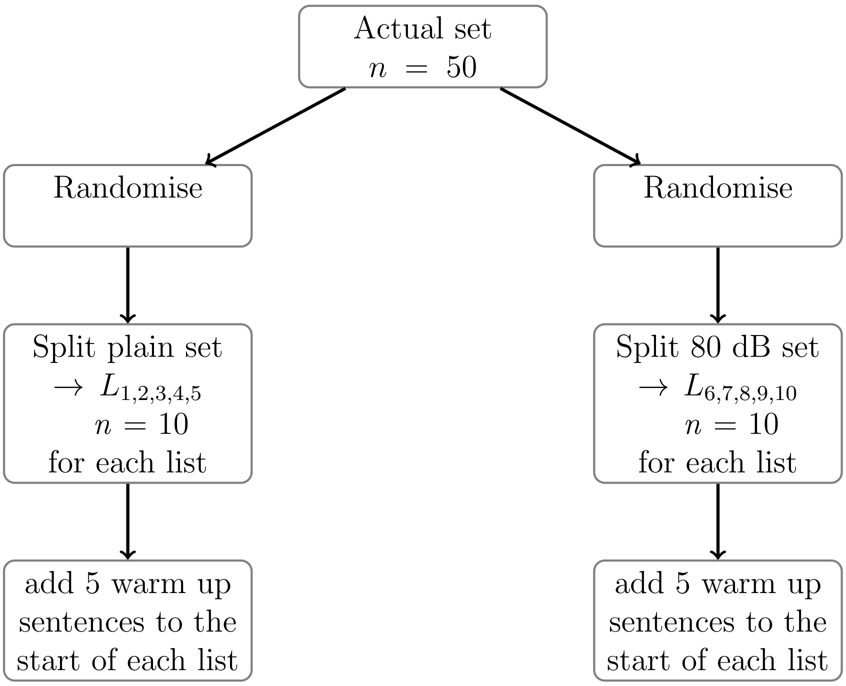


Figure 7.4: The preparation of the prompt lists.

the Lombard recording at 80 dB. This generated 10 prompt lists that were artificially balanced: 5 plain lists and 5 Lombard lists. The warm-up set was split into 10 unique subsets; each fed a prompt list with 5 warm-up sentences. In total, a prompt list contained 15 sentences: 5 warm-up sentences followed by 10 actual sentences.

A talker read one prompt list in a session. A session could be a plain session (with no masker presented to the talker), in which the talker read a plain prompt list, or a Lombard session (with the masker presented), where the talker read a Lombard prompt list. The recording was done in 2 blocks of 5 sessions (10 sessions in total: 5 plain and 5 Lombard). The order of the sessions was governed by the recording schedule that was randomly assigned to each talker. Two example recording schedules are shown in Table 7.1.

The audio and the video recording followed the same procedure described in Section 7.2.3. Figure 7.5 illustrates the collection setting. The talker's task was to read the sentences to the researcher who acted as a listening partner. Having a listening partner in the recording setup was necessary because the Lombard effect is not only triggered as an unconscious reaction to noise, but also by the need to maintain intelligible communication in noise [193].

The talkers were seated inside the booth facing Screen 1 in Figure 7.5 on which the prompt lists were presented, and the listener was seated outside the booth. Based

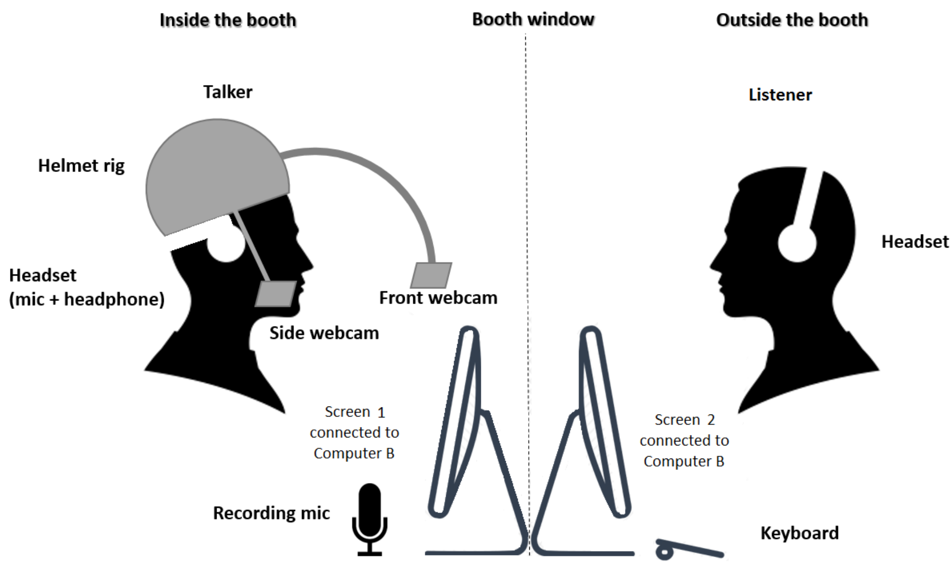


Figure 7.5: The recording procedure.

on the results of the pilot study presented in Appendix A.3, which found a possible psychological effect on the talker due to being able to see the listener, face-to-face interactions were prohibited by placing a white paper screen on the window separating the talker from the listener. The listener listened to the talker's speech presented at 60 dB SPL via a pair of Panasonic RP HT225 headphones connected to the audio interface. The presentation of the sentences and the listener's messages to the talker were controlled by a MATLAB script running on computer B (see Section 7.2.3). The script controlled two interfaces: one for the talker which presented sentences and the listener's messages (Screen 1) and one for the listener which controlled the presentation of the sentences/messages (Screen 2).

The talkers were instructed to talk at a normal pace and in a natural style, and were given 5s to read each sentence. To aid this process, the talkers were prompted by a progress bar on Screen 1 with a duration of 5s. If the talker misread the prompt, the listener presented the same sentence again. In the Lombard sessions, the listener asked the talker to repeat an utterance every 5 to 7 sentences by indicating that she could not hear the talker. The purpose of this step was to maintain the public Lombard loop which is driven by the communication need [176].

Overall, the talkers uttered a collection of 8,250[7] warm-up and actual sentences; these included 4,125 Lombard sentence and 4,125 plain references.

[7]In the final production stage, the number of the collected utterances exceeded 8,250 as they also included repeated and misread utterances.

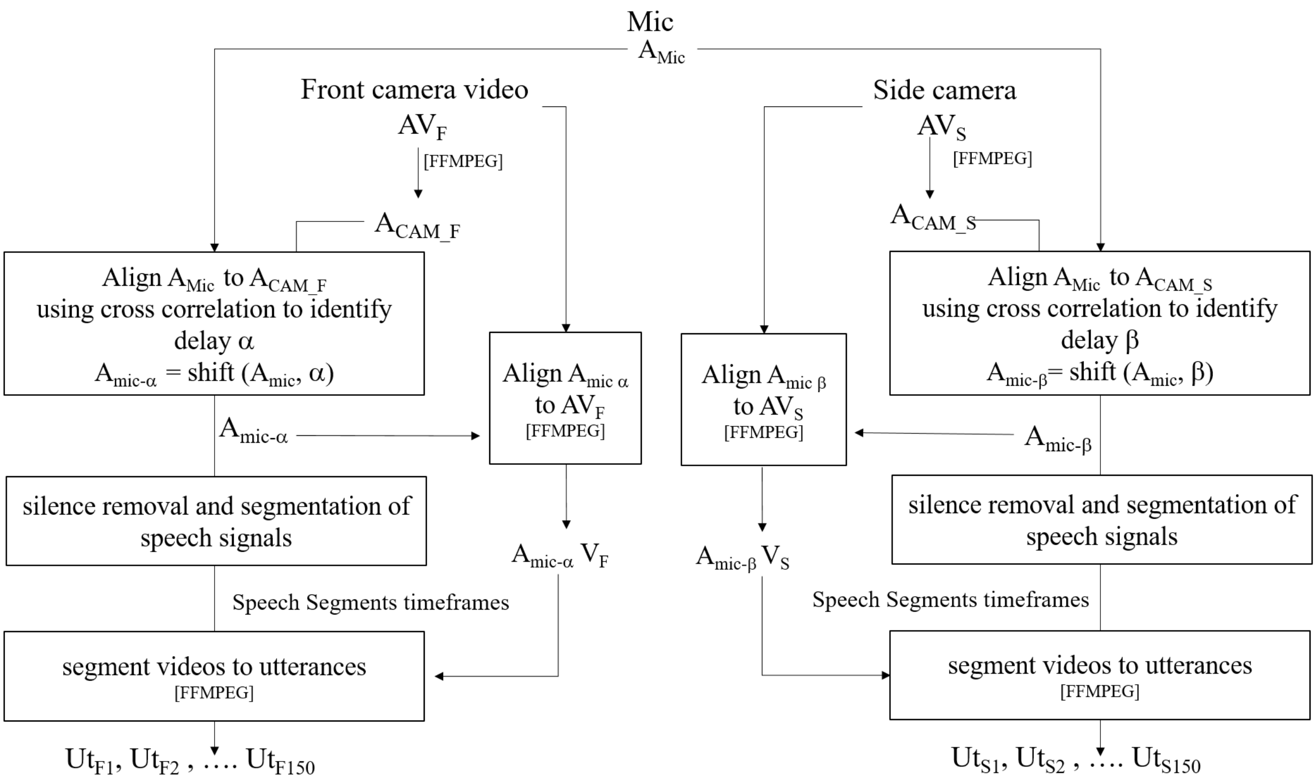


Figure 7.6: The segmentation framework for the recording data used in the analysis. Symbols: A: audio, AV: video, Ut: utterance. Subscripts: Mic: audio source is the microphone, F: video source is front camera, S: video source is the side camera, CAM_F: audio source is the front camera, CAM_S: audio source is the side camera, mic-α: the microphone audio is shifted by α, mic-β: the microphone audio is shifted by β.

7.2.5 Post-processing

The collected raw data was post-processed into a suitable format to facilitate the intended analysis in Section 7.3. Figure 7.6 describes the post-processing procedure. The audio collected from the microphone, A$_{mic}$ in Figure 7.6, was used to guide the segmentation for the videos. The following describes the segmentation steps for the front videos, AV$_F$:

1. Step 1: align the microphone audio with the video audio. The audio of the front video, A$_{CAM_F}$, was extracted using FFMPEG [22] scripts. A$_{mic}$ was aligned with A$_{CAM_F}$ using a cross correlation function [255]:

$$(A_{mic} \star A_{CAM_F})[\alpha] = \sum_{m=-\infty}^{\infty} A^{\star}_{mic}[m] A_{CAM_F}[m+\alpha] \qquad (7.1)$$

where A_{mic} and A_{CAM_F} are the input signals, α is the lag (displacement) between the input signals, and $A^{\star}_{mic}$ is the complex conjugate of A_{mic}. This function was implemented by calculating the product of the Fourier transform of A_{CAM_F} and the conjugated Fourier transform of A_{mic}. The lag α is then identified as the maximum of the cross correlation output in which the two signals are best aligned. A_{mic} is shifted by the value of α to produce $\mathrm{A}_{mic,\alpha}$ which replaces A_{CAM_F} in AV_F creating the video sequence $\mathrm{A}_{mic,\alpha}\mathrm{V}_F$.

2. Step 2: segment $\mathrm{A}_{mic,\alpha}\mathrm{V}_F$ into utterances. Segmentation was achieved by thresholding the signal energy and spectral centroid extracted from an amplified version of $\mathrm{A}_{mic,\alpha}$ to detect speech segments in $\mathrm{A}_{mic,\alpha}$ [109, 110] (Figure 7.7). The output from this process defined the onset time for each utterance presented in the input audio track. The segmentation of an utterance started 30s before the onset of the utterance. This 30s margin was included to accommodate anticipatory visual speech cues (this is due to the natural asynchrony between the auditory and the visual speech signals – i.e., the onset of the mouth movement and the onset of the acoustic production of a speech are not aligned). Four seconds is the duration of a segment, which includes the utterance frames, bounded by some silence frames. This segmentation re-encoded the front video using the FFMPEG encoder x264 that created H.264 videos, making the frame rate of the segmented videos a fixed rate (24 fps). All utterances produced in response to the listener's repeat requests were discarded.

A similar process was repeated to segment the side videos AV_s, creating synchronised front and side utterances. The profile video encoding is similar to the original raw data (encoded using WMV2). The segmentation also copied the original encoding of the audio stream in the input video. All segmented videos are of 4s duration; each front video has 96 frames, while each side video has 120 frames. The analysis in this chapter only considers the front videos.

In summary, and after discarding warm-up sentences, the speech materials in this dataset consist of 5,500 segmented full-face videos, 5,500 segmented profile videos and 5,500 segmented audio signals each representing a single sentence. There are 2,750 unique utterances spoken in a Lombard condition and 2,750 corresponding non-Lombard reference utterances (i.e., the same sentence spoken by the same speaker).

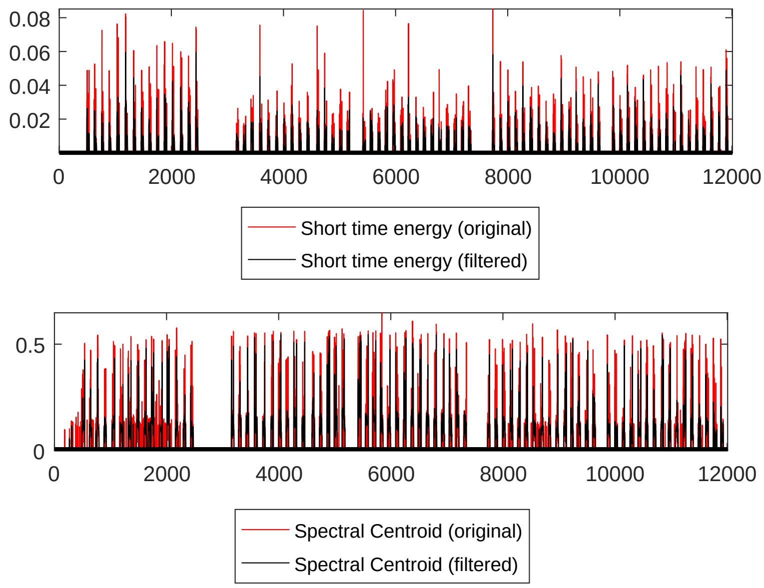


Figure 7.7: Filtering energy and spectral centroids by using thresholds to detect speech segments.

7.3 Phonetic Analysis for Visual Lombard Speech

In this section, the impact of the visual Lombard effect at the phoneme level is examined. This is done by characterising the change in phonemes' visual spaces using a function of visual articulatory geometric measurements associated with the production of each phoneme.

7.3.1 Speech Corpus

Lombard utterances and their reference plain sentences taken from four male (IDs: S14, S46, S47, and S55) and four female (IDs: S7, S21, S32, and S44) talkers, who were chosen at random, were used to provide the pool of phoneme frames for analysis. Each talker's data consisted of 25 pairs of sentences in plain and Lombard conditions; 50 utterances in total. All letters in the alphabet (except for W) were present in the selected sentences for each talker to give phonetic variation in the pool; 31% of speech sounds were vowels and 69% were consonants. The plain video utterances for each talker were concatenated using an FFMPEG script into one video track; a similar process was also done to the Lombard video utterances. The purpose of this step was to facilitate the facial landmark annotation and training process in the Faceware Analyser (FA) tool (see Section 5.2). Word and Phone level transcriptions

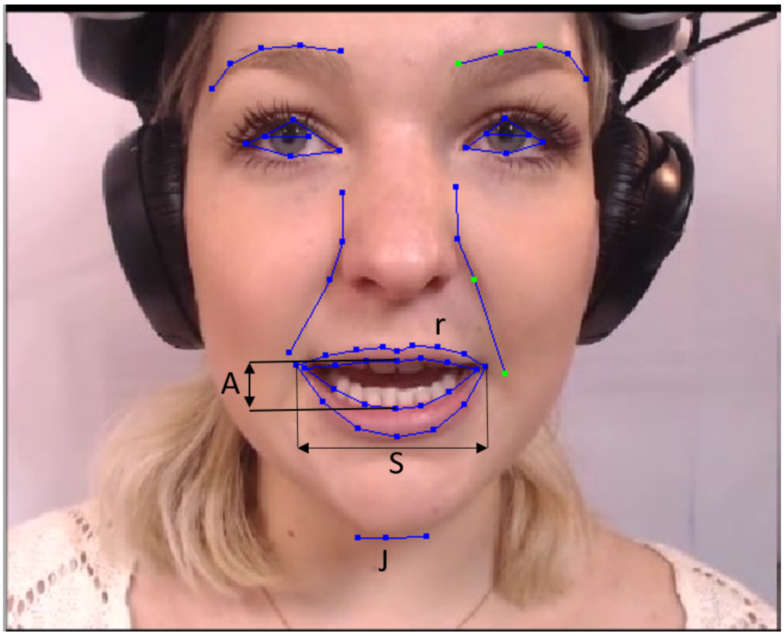


Figure 7.8: Visual articulatory features: A for lip aperture; S for lip spreading; r for lip rounding and J for jaw vertical position.

of the audio tracks of the combined videos were extracted using a web-based set-up of the Penn Phonetics Lab Forced Aligner [171], which is a python-based interface to Hidden Markov Model Toolkit (HTK) [338]. Phonemes in the forced alignment were presented in Arpabet notation – Table 7.2 maps between Arpabet and IPA notations. The transcriptions were manually refined using Pratt [32], and were aligned to their associated video frames using a MATLAB script. The same script labels each frame that falls within a phoneme boundary into either an onset frame (1^{st} frame), a final frame (n^{th} frame where n = number of frames in a phoneme instance), or a mid frame (i^{th} frame where $1 < i < n$). All silence frames were excluded from the pool. About 31% of the speech sounds in this pool are vowels Figure 7.9 illustrates the duration (in video frames) of phonemes in the analysis pool. Consistent with the acoustic analysis of phonemes reported in [193], all phonemes are characterised by longer duration under Lombard conditions. To facilitate the acoustic analysis, audio-only utterances in the same condition of a talker were end-pointed to remove silence frames, and the resulting utterances were combined using an FFMPEG script into one track (i.e every talker was associated with one plain audio track and one Lombard audio track).

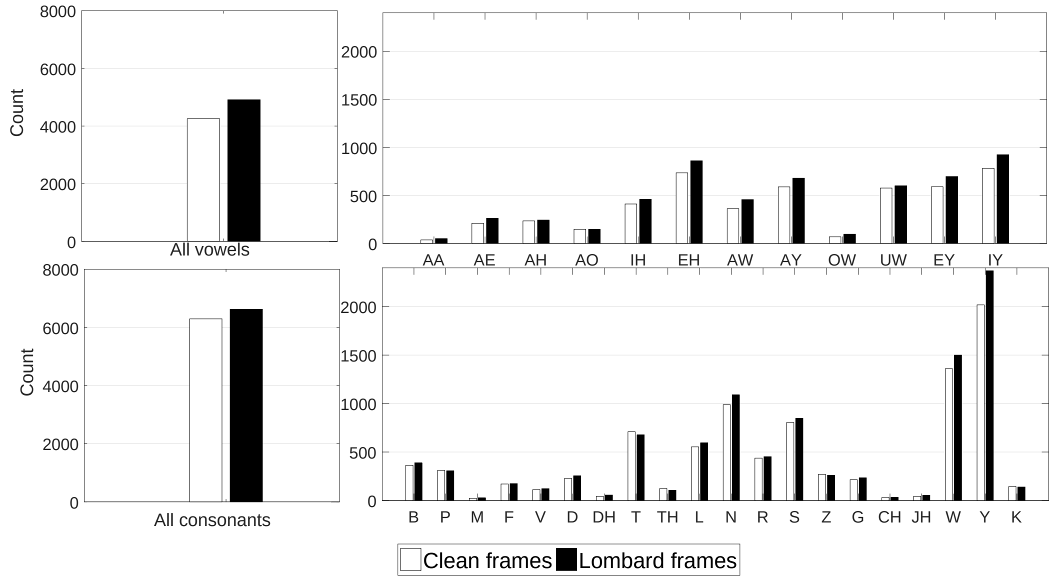


Figure 7.9: The number of frames in plain and in Lombard utterances. Upper row: vowels; Bottom row: consonants; Left column: phoneme category; Right column: individual phonemes.

7.3.2 Acoustic and Articulatory Features

Two acoustic features were estimated from the audio-only utterances tracks. Using Praat [32], the average power magnitude (RMS) of all the samples and the average of the valid F0 estimates were calculated. Four geometric articulatory measurements used in previous literature [101–103, 306] were calculated from facial landmarks extracted from videos using FA (Figure 7.8). These included the following:

- Lip horizontal aperture, or the spreading (S), which is the horizontal distance between the right and left lip corners (mouth landmarks 4 and 10 in Figure 5.1).

- Lip vertical aperture (A), which is the vertical distance between the top and the bottom middle of the inner mouth contour (mouth landmarks 25 and 19 in Figure 5.1).

- Lip rounding (R), which is obtained by $R = 1 - e$, where e is the eccentricity of an ellipse fitted to the outer lip contour (mouth landmarks 1 to 14 in Figure 5.1) given by $e = \sqrt{1 - \frac{a^2}{b^2}}$, where a and b are the lengths of the semi-major axis and the semi-minor axis of the ellipse, respectively, and $0 \leq e \leq 1$. The eccentricity of a circle $= 0$, therefore, the closer the value of e to 0, the rounder the shape of the mouth.

Arpaebt	IPA	Arpabet	IPA	Arpabet	IPA	Arpabet	IPA
AO	ɔ	AW	aʊ	B	b	L	l
AE	ɑ	AY	aɪ	F	f	N	n
AH	ʌ	EY	eɪ	M	m	R	r
EH	ɛ	OW	oʊ	P	p	S	s
IH	ɪ	W	w	V	v	Z	z
IY	i	Y	j	T	t	TH	θ
UW	u			D	d		

Table 7.2: Arpabet notation vs. IPA notation.

- The vertical jaw location (J) which is given by the y-value of jaw landmark 2 in Figure 5.1.

To extract the associated facial landmarks, a talker-dependent FA tracker was trained using a training set which included 70 manually annotated mouth (26 points each) and jaw (3 points each) shapes. For some talkers, the helmet may have slightly displaced backwards as a result of the talkers' movement during the recording, which could affect the camera-talker distance. To correct for this, all landmarks were divided by the Euclidean distance between the midpoint of the inner corners of the eyes and the point making the tip of the nose, which are not affected by articulation. All visual articulatory features for a talker were normalised by their corresponding minimum and maximum mouth movements that talker made in the recording (Table 7.3). Based on this, the visual articulatory measurements are on a [0 - 1] scale.

7.3.3 Utterance Level Analysis

Acoustic Analysis

Figure 7.10 presents a summary of the acoustic analysis conducted on the power amplitude and F0. Consistent with previous research [152, 193], a significant difference in RMS energy and in mean F0 between plain and Lombard conditions is noted as shown by the non-overlapping standard error bars (Figures 7.10a, b, f and g). A related-sample t-test suggests a significant difference in RMS energy (t = 7.8, p = 0.0005) and in mean F0 (t = 7.5, p = 0.0006) between plain and Lombard conditions. A shift towards higher energy bands in the Lombard power amplitude histogram (Figure 7.10c) is observed with a similar shape to the plain histogram. The Lombard F0 data histogram (Figure 7.10h) features a flattened and skewed histogram towards high frequency bands.

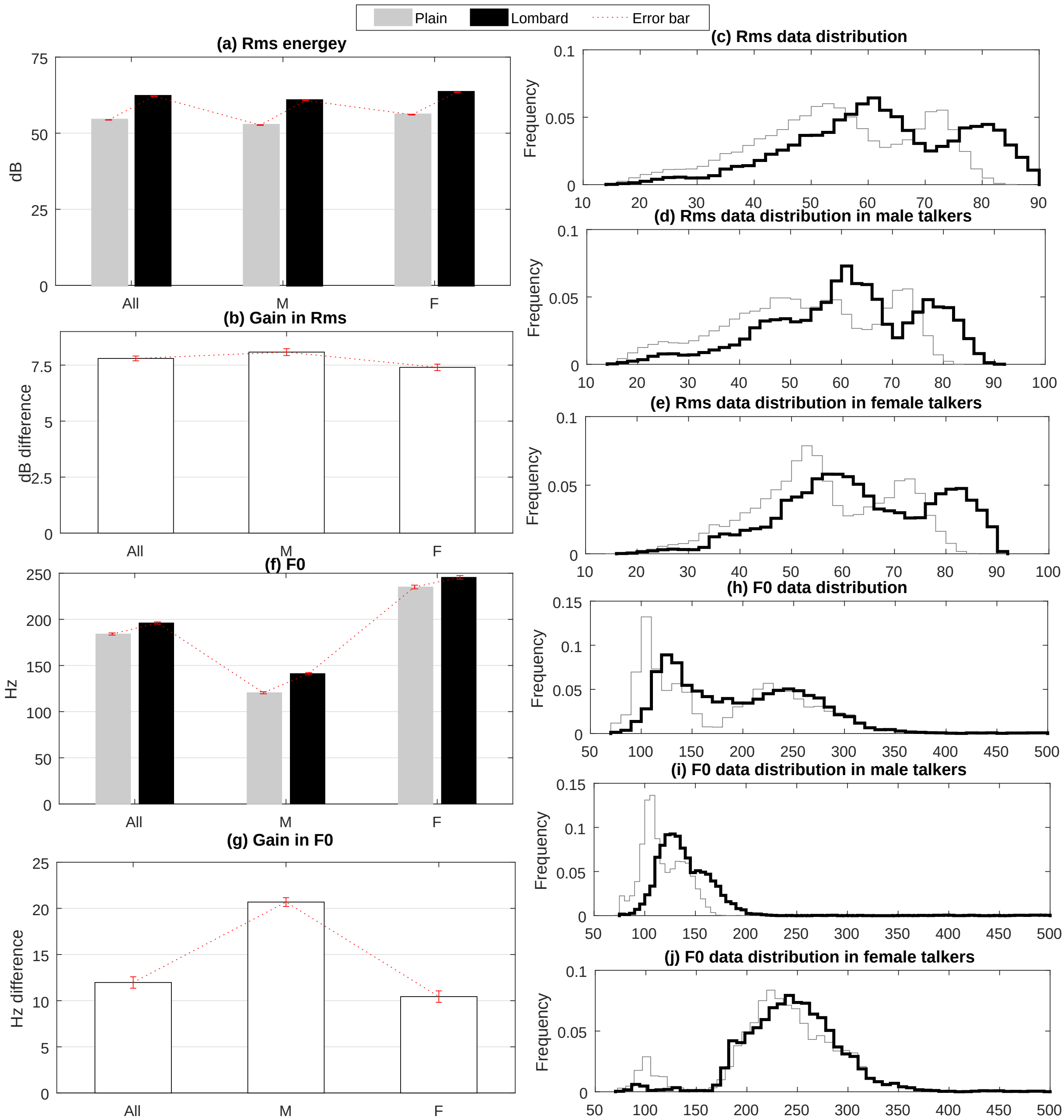


Figure 7.10: Acoustic analysis across all talkers (All), all male talkers (M) and all female talkers (F): (a) RMS energy (b) gain in RMS energy under Lombard conditions. Histogram of power amplitude taken from (c) all talkers, (d) male talkers, (e) female talkers. (f) F0 data, (g) gain in F0 under Lombard conditions. Histogram of F0 data taken from (h) all talkers, (i) male talkers, (j) female talkers.

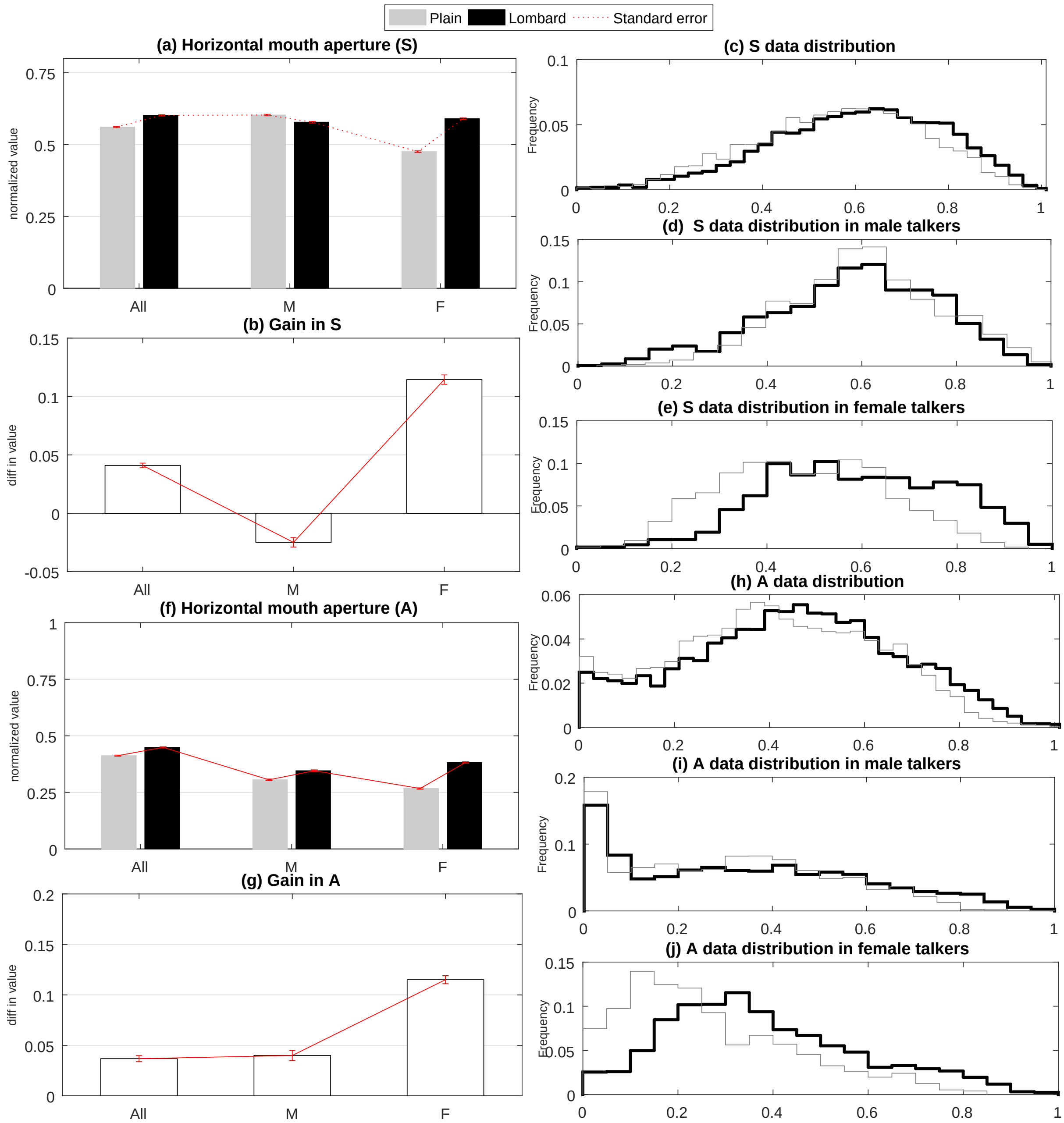


Figure 7.11: Part 1: visual articulatory features across all talkers, all male talkers and all female talkers: (a) horizontal mouth aperture (S) (b) gain in S under Lombard conditions. Histogram of S taken from (c) all talkers, (d) male talkers, (e) female talkers. (f) vertical mouth aperture (A), (g) gain in A under Lombard condition. Histogram of A data taken from (h) all talkers, (i) male talkers, (j) female talkers.

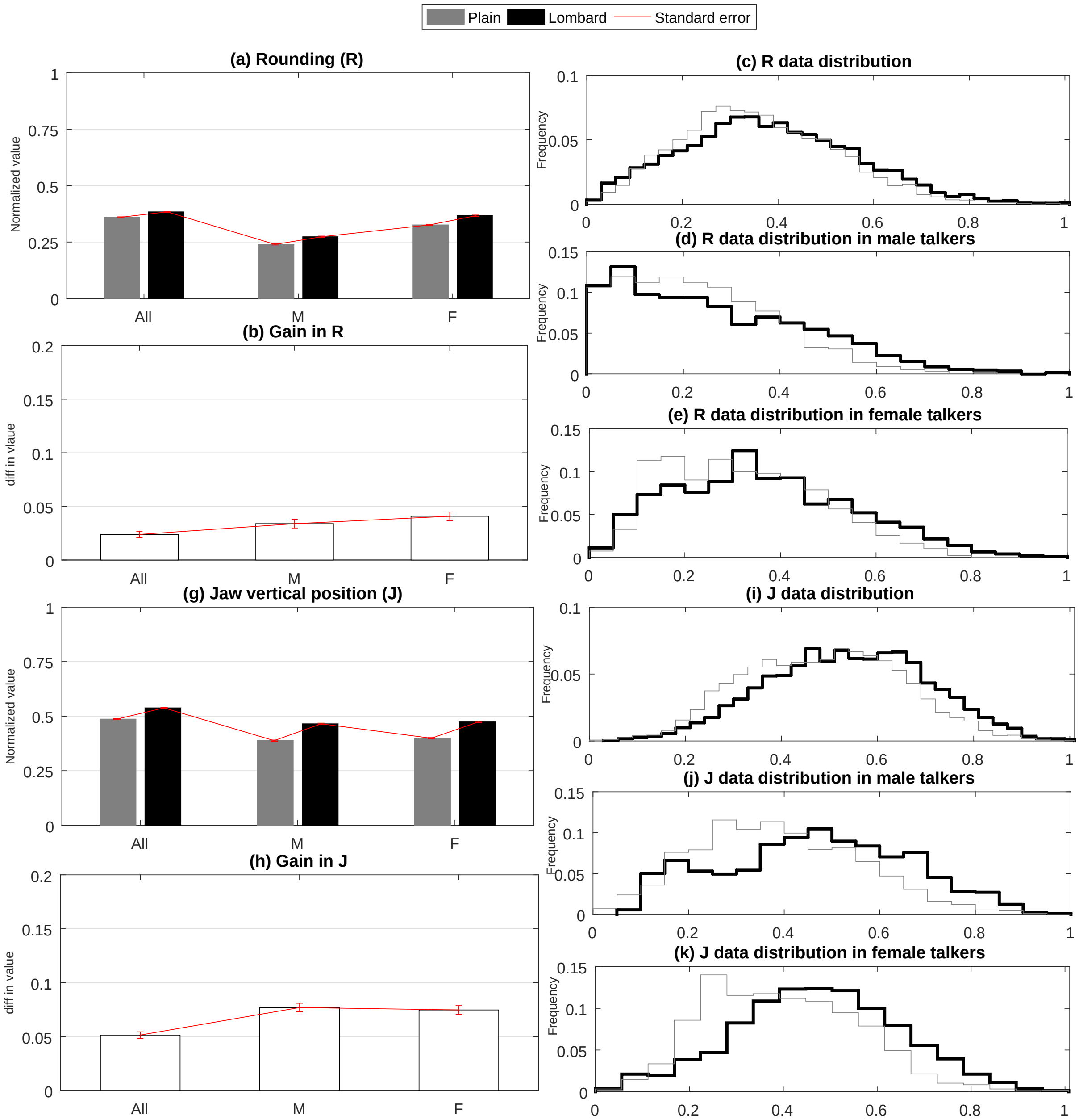


Figure 7.12: Part 2: visual articulatory features across all talkers, all male talkers and all female talkers: (a) rounding (R) (b) gain in R under Lombard conditions. Histogram of R taken from (c) all talkers, (d) male talkers, (e) female talkers. (f) vertical jaw position (J), (g) gain in J under Lombard condition. Histogram of J data taken from (h) all talkers, (i) male talkers, (j) female talkers.

	S		A		R		J	
	min	max	min	max	min	max	min	max
S7	266.10	324.86	0	66.81	0	.057	405.69	446.3
S14	268.32	338.86	0	47.52	0	.036	419.76	450.33
S21	277.03	348.60	0	48.62	0	.032	414.96	435.40
S32	82.352	116.65	0	43.10	0	.41	373.24	398.52
S44	236.89	312.73	0	44.16	0	.04	384.96	416.92
S46	122.46	135.95	0	37.68	0	.21	404.01	427.15
S47	268.32	344.16	0	27.36	0	.02	398.44	429.50
S55	111.02	168.57	0	46.75	0	.24	419.23	442.94

Table 7.3: In pixels, the minimum and the maximum gestures each talker made in the recording. These values were used to normalise the articulatory measurements for each talker.

A gender-difference in acoustic change under Lombard conditions is also noted, a similar finding to Junqua [152]. Figure 7.10a shows that female talkers are characterised with a higher baseline plain RMS energy than male talkers; the mean F0 is always higher for female talkers due to the difference in biological structure of the vocal cords in both genders [285]. Male talkers, however, show a relatively higher gain in RMS energy (diff = 0.68 dB) and a significantly higher mean F0 (diff = 10.2 Hz, 2 semitones). A similar data behaviour reflected by the double-peaks shape of the histograms in Figures 7.10d and e is observed in male and female talkers, with a slight shift towards higher energy bands in male talkers under Lombard conditions. The histograms of F0 data (Figure 7.10i and j) in male and female talkers are different as a result of the classic gender difference in F0 values. They are both, however, skewed towards higher values under Lombard conditions.

Articulatory Analysis

Figures 7.11 and 7.12 provide a summary of the articulatory analysis conducted on the horizontal mouth aperture (spreading)(S), vertical mouth aperture (A), mouth rounding (R), and vertical jaw position (J). Consistent with previous research [63, 91, 100], a significant difference in these measures between plain and Lombard condition is noted as reflected by the non-overlapping standard error bars (Figures 7.11a and f and Figures 7.12a and f). A related-sample t-test suggests that there is a significant difference in these measures between plain and Lombard conditions across all talkers. For example, a related-sample t-test reported a significant difference in mouth opening ($t = 6.29$, $p = 0.0004$). The histograms of the articulatory measures (Figure 7.11c and

Monophthong vowels	AE, AH, AO, EH, IH, IY, UW
Diphthong vowels and Semi vowels	AW, AY, EY, OW, W, Y
Labial consonants	B, F, M, P, V
Coronal consonants	D, L, N, R, T, TH, S, Z

Table 7.4: Phoneme categories.

h and Figure 7.12c and h) show a shift in the Lombard histograms as data migrates to higher energy bands.

A gender difference in articulatory change under the Lombard condition is observed. Figures 7.11a and f suggest that male talkers (in this sample) are more articulate than females in the baseline plain conditions as shown by the magnitude of mouth opening and spreading. Female talkers seem to produce more energetic rounding cues in the baseline condition than male talkers (Figure 7.12a). No significant difference is noted between male and female talkers in jaw energy in plain conditions (Figure 7.12f). This is consistent with Tang *et al.*'s work [306] which found greater visual speech modifications by male talkers in plain and clear speech than by female talkers. In Lombard conditions, however, female talkers made greater modifications in S, A, and R magnitudes, and therefore produced more pronounced speech than male talkers. In fact, male talkers produced less spreading movement in the Lombard conditions than in the plain conditions. Moreover, both the male and female talkers made comparable energy gains in jaw movements. Figures (7.11 – 7.12)d, e, i and j illustrate a shift to higher values in the histograms of the male and female talkers in all articulatory measures (except for the S data in the male talkers).

7.3.4 Phoneme-level Analysis

Figure 7.13 shows the global change in articulatory measures for vowels and consonants by illustrating the data histograms in the plain and Lombard conditions. A shift towards higher energy bands in the Lombard histogram data for vowels was observed in all measures (Figures 7.13a, b, c and d). Consonants featured a similar shift in mouth spreading and jaw movement (Figures 7.13e and h), but there was no prominent change in mouth opening and rounding data (Figures 7.13f and g) under the Lombard conditions.

Interactive visual analytics software was developed to visualise the articulatory measures at the phoneme level. The software mapped the phoneme stream with their articulatory measures using the phonetic alignment of the utterances' audio tracks and the facial landmarks XML file extracted from the utterance videos (see Appendix

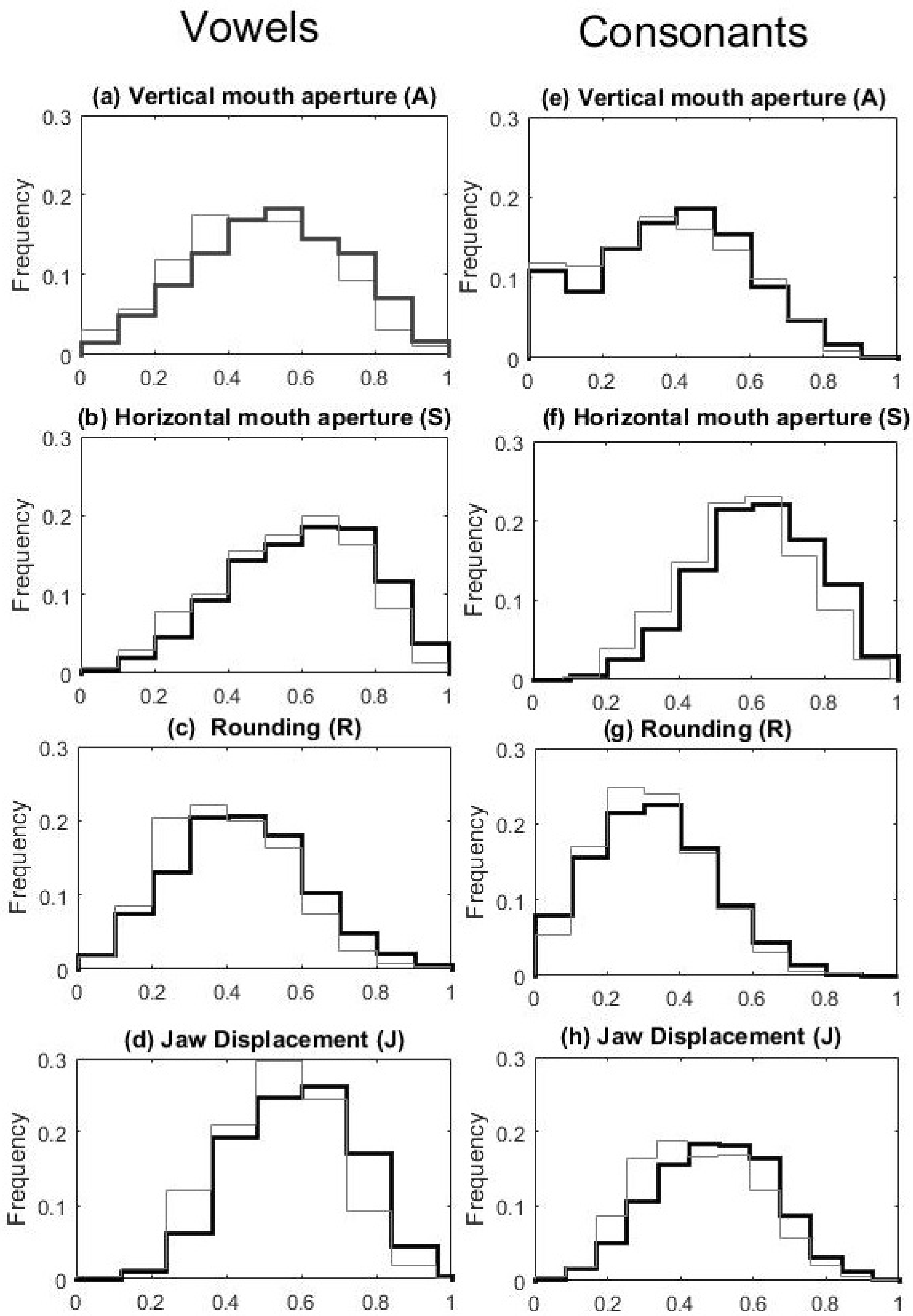


Figure 7.13: Global modification in articulatory measures in vowels and consonants under Lombard condition. Grey histogram: plain, black histogram: Lombard.

A.4 for more information about the design of the software). To simplify the analysis, phonemes were grouped into four categories based on their place of articulation (Table 7.4). All occurrences of all phonemes in these categories were considered in this analysis, and each articulatory measure of a phoneme occurrence was equal to the mean of that articulatory measure taken from the three centre frames.

Figure 7.14 shows the articulatory measures of monophthongs (the remaining figures for each category can be found at Appendix A.4 (Figures A.8-A.11). In both the plain and Lombard conditions, variability in articulatory measures is observed for each phoneme (Figures A.8-A.11); this can be seen in the differences between the minimum and maximum values. This variability can be attributed to a number of things. First, these phonemes were extracted from different contexts (words), which increases the variation due to the co-articulation effect. A second aspect is related to the hypo- and hyper-articulation (H&H) theory [187], in which the energy of speech production fluctuates from a hypo to a hyper style over time. This can make some contexts more energetic than others. Such variability, however, was observed less in the phonemes in the Lombard conditions: the plotted points of the articulatory measures in Figures A.8-A.11 become closer to each other and created small clusters that featured an equal or a very comparable value, and in some phonemes, all points converge, reducing the standard deviation and hence the variation among these points.

Table 7.5 summarises Figures A.8-A.11 by characterising the articulatory modifications in phonemes under the Lombard conditions using five different indices, four of which were generated by subtracting the means in the plain conditions (P) from the means in the Lombard conditions (L):

- the spreading index, $\Delta S = \overline{S_L} - \overline{S_P}$;
- the opening index, $\Delta A = \overline{A_L} - \overline{A_P}$;
- the rounding index, $\Delta R = \overline{R_L} - \overline{R_P}$;
- the jaw index, $\Delta J = \overline{J_L} - \overline{J_P}$;
- and the fifth is the hyper-articulation index, $HI = |\Delta S| + |\Delta A| + |\Delta R| + |\Delta J|$.

Using these indices, Figure 7.15 orders the phonemes from the least energetic to the most energetic under the Lombard conditions. As expected, all vowels were characterised as having the most energetic behaviour under the Lombard conditions. The majority of monophthongs and diphthongs featured the highest modifications

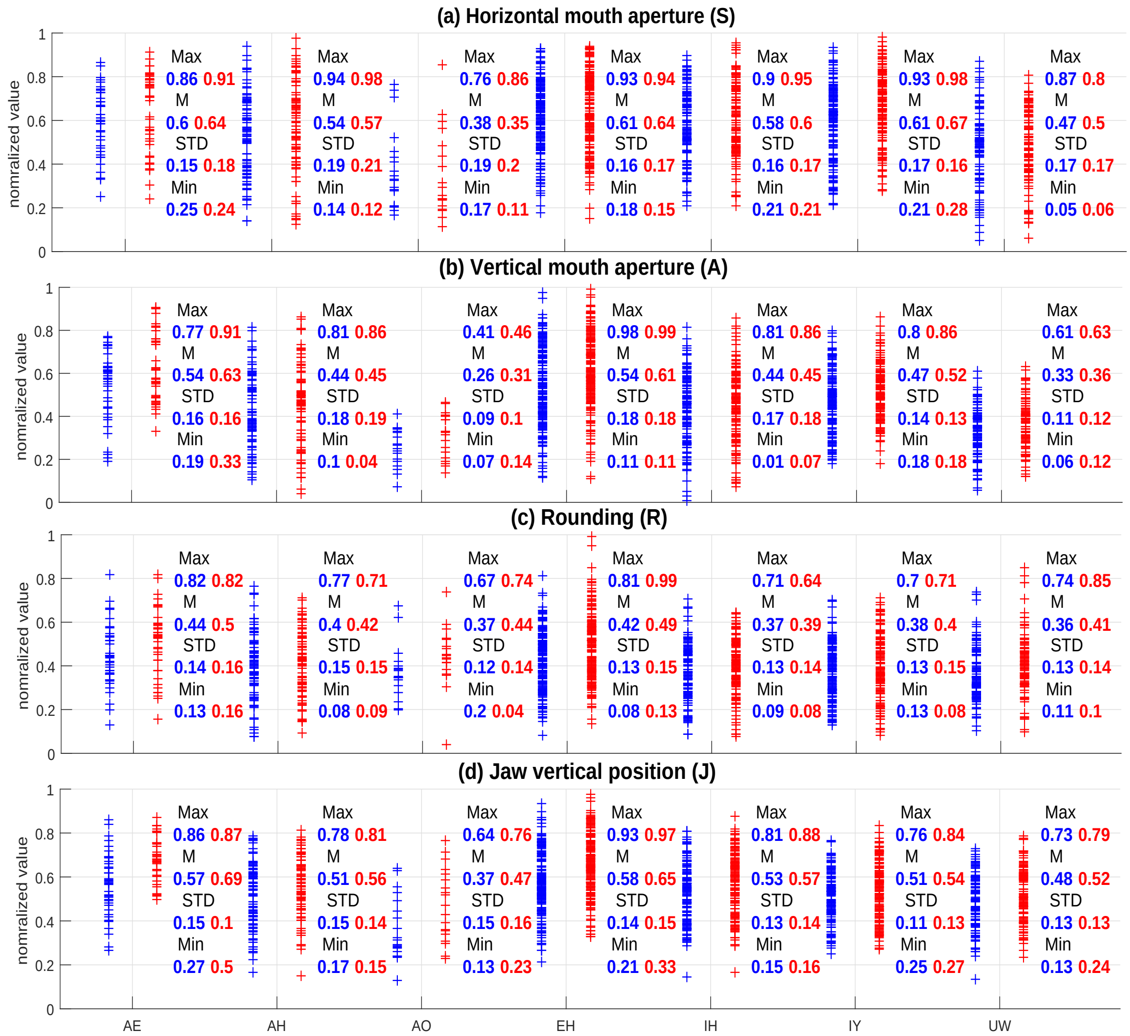


Figure 7.14: Articulatory modifications in monophthong vowels taken from all talkers. (a) horizontal mouth aperture; (b) vertical mouth aperture (c) rounding (d) jaw displacement. Blue: plain, red: Lombard.

Phone	ΔS	ΔA	ΔR	ΔJ	**HI**	**Phone**	ΔS	ΔA	ΔR	ΔJ	**HI**
Monophthong vowels											
AE	0.04	0.09	0.06	0.12	**0.31**	**EH**	0.03	0.07	0.07	0.07	**0.24**
AH	0.03	0.01	0.05	0.04	**0.13**	**IH**	0.02	0.01	0.02	0.04	**0.09**
AO	-0.03	0.05	0.07	0.10	**0.25**	**UW**	0.03	0.03	0.05	0.04	**0.15**
IY	0.06	0.05	0.02	0.03	**0.16**						
Diphthong vowels and Semi vowels											
AW	0.01	0.09	0.05	0.10	**0.25**	**OW**	0.01	0.08	0.07	0.09	**0.25**
AY	0.02	0.05	0.05	0.06	**0.18**	**W**	0.01	0	0.01	0.05	**0.07**
EY	0.06	0.06	0.04	0.03	**0.19**	**Y**	0.01	0.04	0.06	0.03	**0.14**
Labial consonants											
B	0.06	0.01	-0.03	0.09	**0.19**	**V**	0.03	0	0	0.06	**0.09**
F	0.01	-0.01	-0.03	0.08	**0.13**	**P**	0.07	0	-0.01	0.05	**0.13**
M	0.07	0	-0.01	0.05	**0.13**						
Coronal consonants											
D	0.04	0.02	0.01	0.03	**0.10**	**T**	0.07	0.04	0.03	0.04	**0.18**
L	0.06	0.01	0.03	0.05	**0.15**	**TH**	0.04	0.03	0.04	0.06	**0.17**
N	0.04	0.04	0	0.03	**0.11**	**S**	0.06	0.04	0.01	0.04	**0.015**
R	0.07	0.04	0.03	0.05	**0.19**	**Z**	0.04	0.03	-0.01	0.01	**0.07**

Table 7.5: A summary of the articulatory modifications for phonemes under Lombard conditions. Five indices that characterise the change: spreading index (ΔS), the opening index (ΔA), rounding index (ΔR), jaw index (ΔJ) and hyper-articulation index (HI).

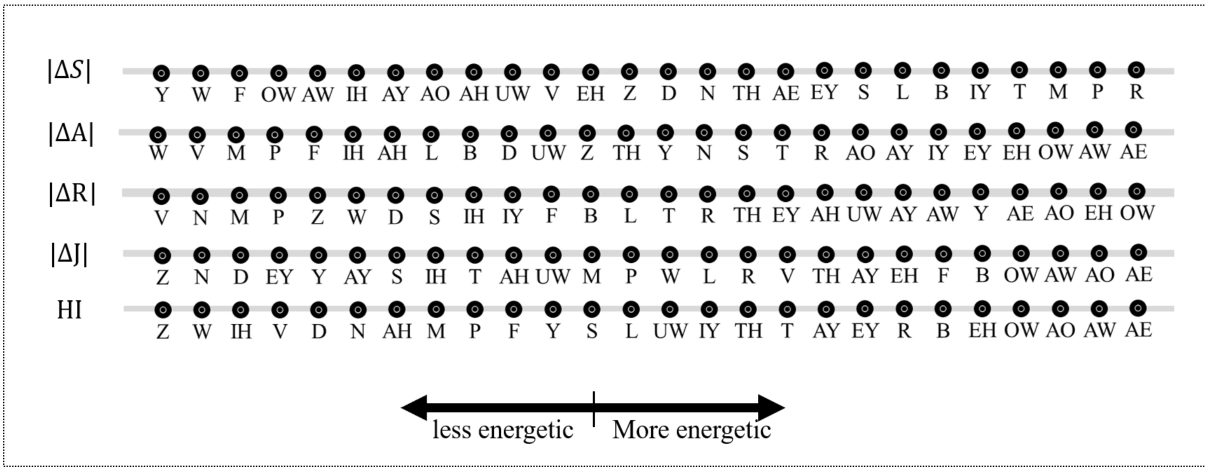


Figure 7.15: Phonemes on a visual energy scale using the absolute values of spreading index (ΔS), opening index (ΔA), rounding index (ΔR), jaw index (ΔJ) and hyper-articulation index (HI).

in mouth opening, mouth rounding and jaw movement. Consonants, however, had larger mouth-spreading modifications. Labial consonants had no change or decrease in the opening and rounding of the mouth, but they did increase the mouth spreading and the jaw's vertical y-position as a result of the increased pressure on the lower-lip under the Lombard conditions. The degree of mouth opening and rounding in the coronal consonants seemed to be sensitive to the coarticulation effect and was more likely to be inherited from a neighbouring vowel. This increase in spreading may have enhanced the visibility of the internal articulators (teeth and tongue), which in turn offered additional cues to facilitate contrasting these sounds.

A more detailed perspective on the phoneme articulatory measures was also considered in this analysis by looking into the plain and Lombard behaviour of phonemes within the same context (i.e., the same word). Letter keywords in the talkers' utterances were selected as the source of the phonemes pool for this analysis. Figure 7.16 illustrates the letter I, which consists of one diphthong (AY), across all talkers. Instead of showing all the figures of the 25 Grid letters, 8 letters were chosen such that their phoneme building blocks could be considered as samples from the phoneme categories illustrated in Table 7.4. The chosen letters are: {E} (monophthongs vowels), {A, I, O} (diphthong vowels); {B} (a labial consonant + a monophthongs vowel); and {T, N, C} (coronal consonants + a monophthongs vowel). The figures of these letters can be found in Appendix A.4 (Figures A.12 - A.19).

There are many observations about the plain and Lombard behaviour within a similar context. First there was a relation between the A, R and J data for all letters. This was expected since phonemes with increased mouth rounding and/or jaw displacement are always accompanied with mouth opening. A negative correlation between S and the $\{A, R, J\}$ data in all vowel letters is also observed. Second, the shape of the phoneme trajectory in the same context is relatively comparable across all talkers. For example, in Figure 7.16b, an arch-shape to the trajectory of the phoneme stream of the letter I across talkers is observed. Individual differences between talkers are also noticed, which explains the more deformed phoneme trajectory shapes in some talkers. One source for these individual differences could be linked to the H&H theory. Talkers seem to be selective in the amount of energy they exert during the production of the phoneme stream; some talkers, for example {S7, S14, S21, S32}, showed a hyper-articulation effect on all phoneme frames, which contributed in preserving the plain trajectory shape in the Lombard data; other talkers, for example {S44, S46, S47} exert the energy at the onset and/or final frames only and, interestingly, show a hypo-articulation effect for the middle frames. Lastly, one talker {S55}, shows a hypo-articulation effect on all frames.

Speaking style may also have an effect on inducing inter-speaker variability, as the talkers seem to give more weight to certain articulatory gestures than others. For example, S44 in Figure 7.16 is observed to give a higher weight to mouth spreading rather than opening across the letters, while S46 does the opposite. The coarticulation effect might also have had an impact on talker variability (to better understand this coarticulation effect, see table A.2 in Appendix A.4, which shows the sentence list from which the letters were extracted).

7.3.5 Discussion

Previous research [63, 100, 102, 103, 163, 164] has addressed the global articulatory modifications in Lombard speech and has found such modifications to be correlated to the vocal effort under the Lombard conditions. The current study has examined the impact of the Lombard effect at the utterance and phoneme levels by characterising the visual spaces of phonemes using visual articulatory geometric measurements. The study has also included an analysis of the impact of Lombard speech by looking at articulatory changes at the phoneme level with consideration for key aspects such as inter-speaker variability, gender and the H&H theory.

Global Modifications

The results regarding global modifications of the acoustic and articulatory properties of Lombard speech are in line with previous research. The impact of gender differences on the global acoustic changes were also consistent with previous research, as the male talkers tend to produce stronger acoustic modifications under the Lombard conditions than the female talkers. The study also found a gender difference impact on articulatory modification in Lombard speech. Surprisingly, the female talkers in this sample tend to produce stronger articulatory modifications than the male talkers. This may suggest that the gender difference impact on Lombard speech is mainly driven by the mechanism of hyper-articulation: male talkers perform acoustic hyper-articulation, while female talkers perform visual hyper-articulation. This can also be considered as evidence that visual modifications under the Lombard conditions are not just correlated to acoustic adaptations and may have an independent communication enhancement goal.

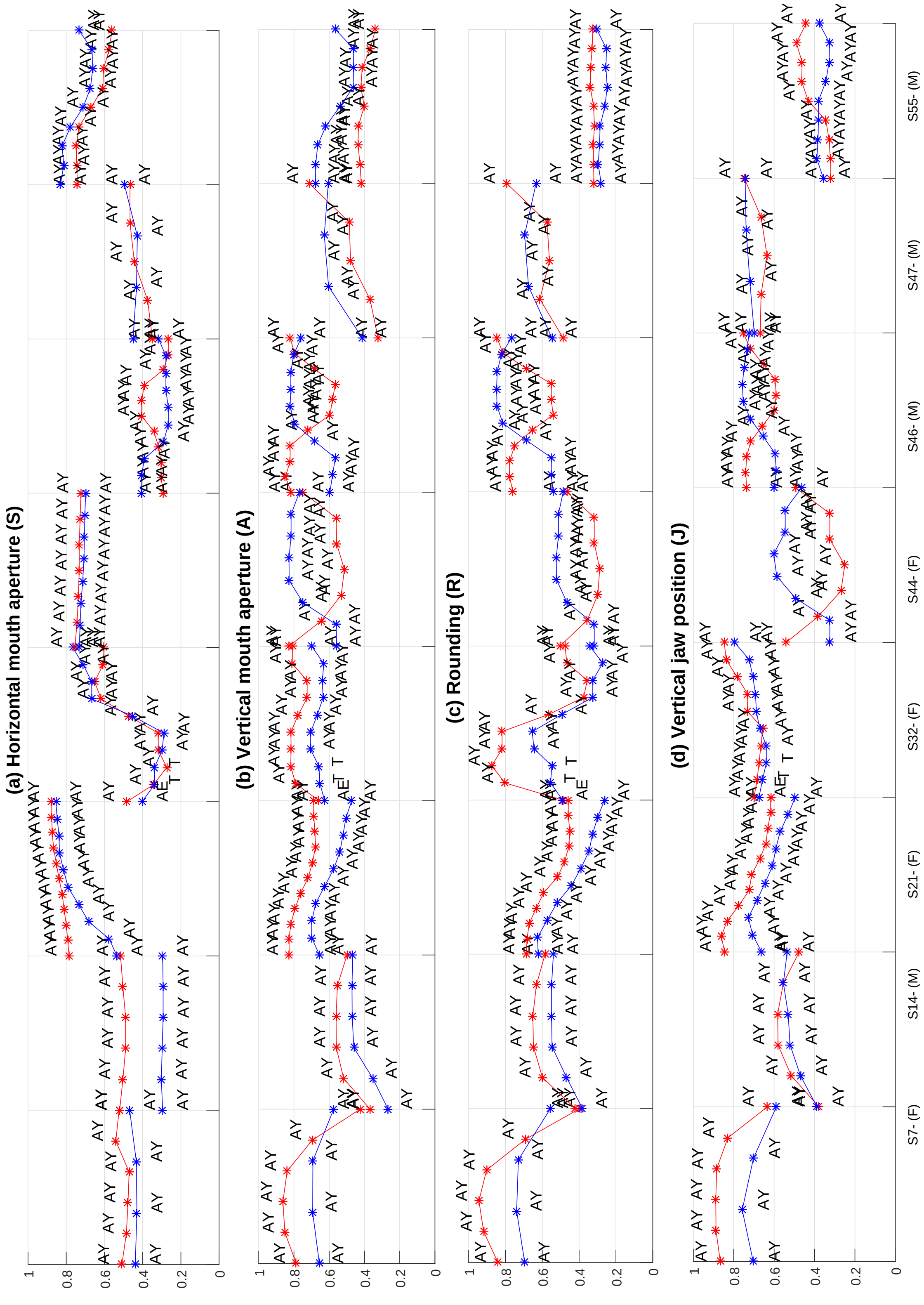


Figure 7.16: The articulatory modifications across talkers when uttering the Letter I in plain and Lombard conditions. (a) horizontal mouth aperture; (b) vertical mouth aperture (c) rounding (d) jaw displacement. Blue: plain, red: Lombard.

Phoneme-level Modifications

The results of the articulatory measures for phonemes reveal that vowels are characterised by stronger global articulatory changes than consonants under the Lombard conditions. This analysis (Figures A.8-A.11) also shows a reduced sparsity in the plotted articulatory data of the Lombard phonemes. This suggests that Lombard phonemes may have some consistent behaviours, such as the way they are produced, and may become less sensitive to factors that induce articulatory variability, such as the coarticulation effect. This might be considered further evidence for the role of articulatory changes under the Lombard conditions for enhancing communication.

By analysing phonemes within a similar context, inter-speaker variability can be seen in articulatory modifications across talkers. H&H theory may explain such variability as each talker behaved differently in their exertion and preservation of articulation energy under the Lombard conditions. This suggests that the impact of H&H theory needs to be carefully considered and modeled, not just acoustically, but also from an articulatory perspective, especially in the field of audiovisual speech recognition.

Articulatory Modification Approximation under the Lombard effect

Given the previous observations regarding the sources of talker variability in visual articulatory modifications in Lombard speech, approximating these articulatory modifications given the articulatory features of plain speech may need to consider the coarticulation effect, the H&H effect, and the speaking style of a talker, or the persona of the talker. Therefore, the articulatory features (Section 7.3.2), $\boldsymbol{l}_{ph\_l_k}$, in the i^{th} frame of the Lombard phoneme ph_l that appeared in the k^{th} context of m distinct contexts [8] of ph_l in Lombard speech corpus, can be expressed as:

$$\boldsymbol{l}_{ph\_l_k}(i) \approx \boldsymbol{p}_{ph\_p_k}(i) + \boldsymbol{am} \tag{7.2}$$

where $\boldsymbol{p}_{ph\_p_k}(i) = \begin{bmatrix} S_{ph\_p_k}(i) & A_{ph\_p_k}(i) & R_{ph\_p_k}(i) & J_{ph\_p_k}(i) \end{bmatrix}^T$ is a set of articulatory features extracted from the i^{th} frame of the plain phoneme ph_p that appeared in the k^{th} context of m distinct contexts of ph_p in the plain speech corpus,

[8] A context is a word in this analysis. For example, phoneme /b/ appeared in four contexts in the collected corpus: 'bin', 'blue', 'by' and 'B'.

and $\boldsymbol{am}$ represents the articulatory modification, which can be expressed as

$$\boldsymbol{am} = \begin{bmatrix} \Delta S_{ph_k} * \boldsymbol{h_S}(i) * sp_S \\ \Delta A_{ph_k} * \boldsymbol{h_A}(i) * sp_A \\ \Delta R_{ph_k} * \boldsymbol{h_R}(i) * sp_R \\ \Delta J_{ph_k} * \boldsymbol{h_J}(i) * sp_J \end{bmatrix} \tag{7.3}$$

$\begin{bmatrix} \Delta S_{ph_k} & \Delta A_{ph_K} & \Delta R_{ph_K} & \Delta J_{ph_K} \end{bmatrix}$ are approximations of the articulatory changes for phoneme ph_l within the k^{th} context. Such approximation can be inferred in a similar way to the process explained in Section 7.3.4 – Table 7.5, however, within the desired context. Such approximation will help to characterise the effect of co-articulation under the Lombard conditions. These articulatory changes can be then scaled (grow or shrink) by the H&H effect, $\boldsymbol{h}$, and the speaking style (talker persona) effect, $\boldsymbol{sp}$.

A possible way to address the H&H behavior of a talker under Lombard conditions is by observing the trajectory paths of the articulatory features of ph_p and ph_l in the k^{th} context in a training set (for example, the trajectory paths constructed from the articulatory features at the Lombard and plain phoneme AY frames in the context "I", Figure 7.16). In this observation, the difference between the articulatory features of ph_p and ph_l frames is calculated and frames with zero-crossing values will be labeled as points of hypo/hyper change. Such data can be used to guide a supervised regression that predicts frames of hypo/hyper change in a given contexts. Based on that, $\boldsymbol{h}(i)$ can be expressed as:

$$\boldsymbol{h}(i) = \begin{cases} -1, & \text{when the nearest zero corssing frame is negative,} \\ 0, & \text{if } i \text{ is a zero corssing frame,} \\ 1, & \text{defalut, or when the nearest zero corssing frame is postive} \end{cases}$$

Addressing the speaking style of a talker can be achieved by assigning a weight for each articulatory gestures made by that talker. To learn articulatory gestures' weights for a given talker, a data reduction techniques, such as the principal component analysis, can be applied to the articulatory features of the talker in the plain recordings. $\boldsymbol{sp}$ can then represent the weight of each articulatory gestures given by their normalised eigenvalues.

Future Work

This analysis has offered increased understanding of the mechanism of hyper-articulated speech. Indeed, hyper-articulation is more sophisticated than just a simple amplification or translation of the phoneme spaces [101, 104]. The automatic exaggeration method presented in Chapter 6 should therefore be revisited using a data-driven method that makes use of the dataset presented in Section 7.2, as well as the analysis findings. This work could include modelling the correlation between the speaking styles in plain and Lombard conditions, the coarticulation effect under the Lombard conditions and the H&H effect on visual speech. Moreover, this dataset could be used for further study of the correlation between acoustic and articulatory changes in Lombard speech in order to answer questions about the extent to which visual speech can be exaggerated without creating an audiovisual conflict (Chapter 6). An alternative path to counter potential audiovisual conflict would be to use the dataset to model the acoustic and phonetic adaptations under the Lombard conditions to exaggerate the auditory speech in addition to the exaggerated visual speech. The modelling of the previous mentioned effects may also feed into a model of automatic audiovisual speech recognition systems under noisy conditions in which these effects are considered.

7.4 Summary

In this chapter, a novel bi-view audiovisual Lombard speech dataset collected under high-SNR level (whereas listeners were exposed to low SNR via headphones) was presented. The dataset, which is an extension of the popular Grid corpus, features two synchronised views of the talker, a front view and a profile view, and offers a plain reference to each Lombard sentence. Initial analysis of this dataset showed prominent acoustic, phonetic and articulatory speech modification in Lombard speech. Acoustic and articulatory phoneme analysis for selected talkers were presented. Gender differences in acoustic and articulatory modification under Lombard conditions were observed. Difference in articulatory energy under Lombard conditions between vowels and consonants, and within consonant categories was also characterised. Variability in articulatory modifications was found when looking into phoneme behaviour within the same context, and hypothesised to be linked to the talker's speaking style, the impact of co-articulation and the theory of H&H.

Chapter 8

Conclusions

8.1 Summary of Thesis

This thesis has investigated visual speech enhancement methods to improve auditory and audiovisual perception. The proposed enhancement methods were tested on non-native normal-hearing listeners. The non-native normal hearing listeners were treated in this thesis as 'proxy' listeners to CI users, i.e., the speech perception chain model of non-native listeners when listening to CI simulated speech was used as a predictor of the performance of the CI users. This is because the potential users of these enhancements are CI users. Using normal hearing listeners in conjunction with CI simulation is an approach used by CI researchers [78, 282] since finding a homogeneous CI user group is difficult due to the variation in CI perception. Non-native listeners were selected, in particular, since they and CI users show similar behaviour in perception as they experience internal adverse conditions and show high sensitivity to visual speech cues when listening to native speech.

Two methods of enhancement have been proposed in this thesis: an appearance based and a kinematics based approach (Figure 8.1); each addresses a defining feature of visual speech: static and kinematics features. The appearance based method modifies the appearance of the talker's lips by applying an automatic lipstick effect that colours the talker's lips in order to increase the saliency of the visual speech. The kinematics method applies an exaggeration effect on the talker's speaking style by amplifying the motion of the mouth. Both methods were tested using the audiovisual training framework introduced in Chapter 4. This was used to test the effect of each enhancement on the listeners' audiovisual perception during the training and the auditory perception after the training. Audiovisual training was used as a platform to test the enhancement following a study that correlates audiovisual training with improved post-training auditory perception [28]. A pilot

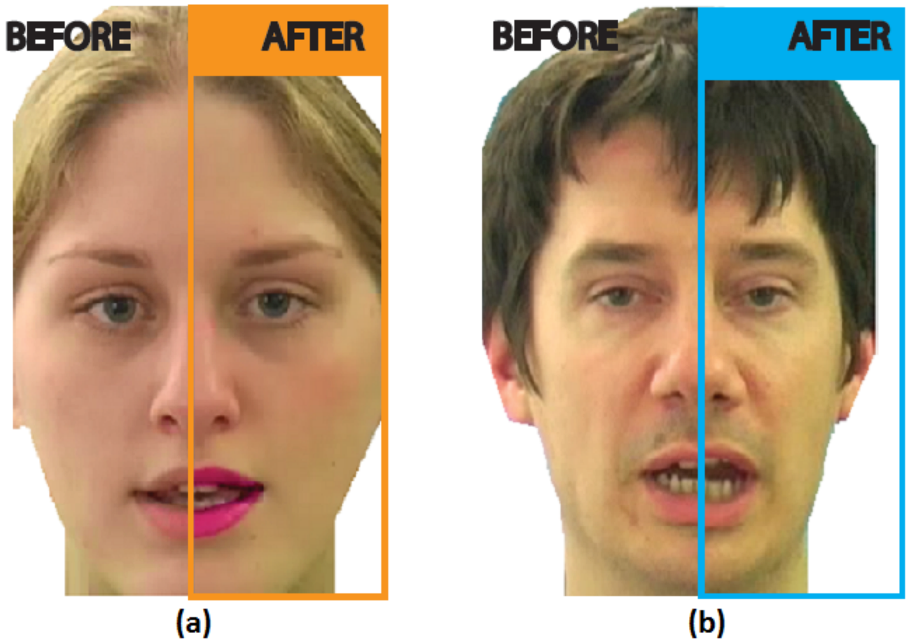


Figure 8.1: (a) the lipstick effect; (b) the exaggeration effect. Each figure shows a talker face before and after applying the effects.

study, introduced in Chapter 4, that evaluated the audiovisual framework suggested the effectiveness of this framework in showing training gains by the listeners.

Chapter 5 covers the appearance based enhancement approach (the lipstick effect). The evaluation produced two sets of results: data from all subjects, and data from selected subjects who showed comparable pre-test abilities. Both sets of results confirmed the lipstick effect's positive impact on improving the auditory and the audiovisual perception of CI simulated speech for the non-native listeners. Such results indicate the usefulness of applying enhancement of visual speech in audiovisual training and encourage investigating more enhancement methods that can modify the static features of visual speech to increase its saliency.

Chapter 6 presented the kinematics-based enhancement approach (the exaggeration effect). The results of using the kinematics-based enhancement in audiovisual training suggest that after exposure to visually exaggerated speech, listeners had the ability to adapt to the conflicting audiovisual signals. In addition, subjects trained with enhanced visual cues achieved better audiovisual perception for a number of phoneme classes than those who were trained with unmodified visual speech. There was no evidence of an improvement in the subsequent audio-only listening skills, however. The subjects' adaptation to the conflicting audiovisual signals may have slowed down auditory perceptual learning and impeded the ability of the visual speech to improve the training gains.

Chapter 7 described the collection and the analysis of a bi-view audiovisual Lombard Grid corpus. This dataset is an extension of the audiovisual Grid corpus [54]. It was collected using a bespoke head mounted camera system. It features two views of the talker in a fixed head-pose, and provides a plain (non-Lombard) reference to each Lombard sentence. The dataset includes 55 talkers that uttered a total of 8,250 utterances: 4,125 Lombard, 4,125 plain utterances. The dataset is processed into a format suitable for analysis.

Chapter 7 also presents an investigation of a real example of kinematics based enhancement in visual speech, by conducting visual Lombard speech analysis of selected talkers from the dataset. The behaviour of visual phonemes in different contexts and across talkers was examined. The analysis has reported a number of findings. First, there was a gender-difference in global modifications in the Lombard visual speech; female talkers produced stronger Lombard speech modifications than male talkers. Second, the visual phoneme behaviour in different contexts seemed to be more consistent and less disparate in the Lombard conditions than in the plain conditions, suggesting a reduced effect of co-articulation in the Lombard conditions. In the same context, variations in visual phoneme behaviour might be linked to the effect of H&H theory [187], in which a talker varies the production energy based on communication demand.

8.2 Future work

Future work would investigate the employment of the enhancement methods, the lipstick and the exaggeration effect, in other training applications such as speech therapy. The lipstick effect can increase the saliency of the talker's mouth shapes. The exaggeration effect can illustrate the combination of key gestures that constitute the mouth shape of a sound, and hence teach trainees to correctly produce that sound. Another possible application is in language training applications. The training profiles (session 1 to session 3) of the non-native subjects in Chapters 4, 5, and 6, show the great potential of such enhancement for improving this group's listening skills.

Another direction for future work is transforming the lipstick effect into a real-time augmented reality solution that can be incorporated into a number of platforms to aid audiovisual perception. Examples of applications that might introduce the lipstick filter include television, YouTube, and video conferencing programs, in which enhancing the visual signal can be analogous to increasing the volume of the auditory signal. Another possible application is to enhance real-life communication by incorporating the lipstick effect into wearable devices used by the listener, such as Google Glass, that track-then-apply the lipstick effect on the interlocutor's lips to enhance the listener's audiovisual perception. Since the current implementation of the exaggeration effect was beneficial in improving the perception of a number of phonemes, it can be introduced alone or combined with the lipstick effect in similar applications when such phonemes are encountered.

Another study could extend the visual Lombard speech analysis conducted in Chapter 7 by including all talkers in the collected dataset in order to examine more phoneme contexts. The bi-view audiovisual Lombard dataset is also expected to serve studies in different fields, such as automatic audiovisual speech recognition, perception studies, computer vision and animation, and behavioural-related studies. It can be used alone, or in conjunction with the Grid corpus, to serve such studies. An example of a study that could use this dataset would model the exaggeration effect by learning from the visual modifications in Lombard speech. Factors that contribute to the variation observed in Lombard speech, mentioned in Chapter 7, would be taken into account, including the talker's speaking style, the effect of co-articulation and the H&H theory. Phonetic and auditory modifications would also be modelled in order to counter the conflict effect observed in visually-exaggerated audiovisual speech.

Appendix A

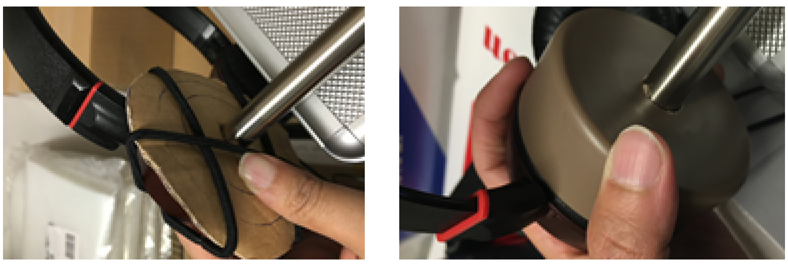
Visual Lombard speech analysis

In this Appendix, further details on some aspects of the dataset recording and the subsequent analysis (Chapter 7) are presented. These include measuring the sound pressure level (Section A.1), and previous prototypes of the recording helmet (Section A.2). A pilot study conducted to inform the dataset collection procedure is also presented in Section A.3. The design of the visual analytics software for visual speech analysis, the Phoneme Viewer, is presented in Section A.4.

A.1 Measuring Sound Pressure level

This section highlights the process of measuring the sound pressure level of the noise masker used in the recording experiment. The sound pressure levels have been measured using a Cirrus Optimus Yellow Class 2. An acoustic coupler, which is needed to seal the headphone during measurement, to avoid noise escape, was not provided with the meter set. So, hand made couplers were used.

The sound pressure level for the room was 26.9 dB SPL. The pair of headphones



(a) (b)

Figure A.1: Acoustic couplers used in the SPL measurements: (a) a cardboard coupler; (b) a plastic coupler.

used for the experiment was a Sennheiser hd 380 pro (used by listener in the dataset recordings), connected to a laptop (Computer B in Section A.3). The volume of the laptop was set to 100% and a MATLAB routine was instead used to control the sound volume. The headphones played a sine wave tone generated at 1kHz. The meter took a 3s recording of the headphones' sound, and the average SPL was computed. The sound volume was reduced using the MATLAB routine until a reading of 80 dB by the SPL meter was recorded. The sound volume was then recorded and used to present the masker in the recording experiment. (detailed in Chapter 7).

Prior to conducting the measurement experiment, two handmade acoustic couplers were tested (Figure A.1): a cardboard and a plastic coupler. On average, when using the cardboard coupler the SPL reading were higher by 4 dB than when using the acoustic coupler. Given these results, the cardboard coupler was used for the subsequent measurement.

A.2 The Recording Helmet

Figure A.2 shows a number of tested prototypes for the helmet used in collecting the recordings in Section 7.2. The aim was to design a lightweight helmet with a stable camera arm that remains fixed during the recording and captures the talker's entire face to facilitate the automatic facial tracking. The early models used a Go-Pro camera that was later replaced with webcams to reduce the weight on the talker's head. Objective evaluation tests were made on the collected videos using these prototypes to check the best talker-camera distance for precise automatic facial tracking. The prototypes were also tested on a number of subjects in order to select the right helmet size for the recording experiment and the maximum tolerable duration a talker to wear the helmet.

A.3 Pilot Study

Prior to collecting the dataset, a pilot study was conducted to examine variables that could regulate the effect of the Lombard speech (see Section 2.5.4 for more information). The pilot study investigated the following:

- The proposed recording duration is three blocks of five sessions. In each session, the talker should read a prompt list of 15 sentences: 5 warm-up followed by 10 actual sentences. The pilot study examined whether or not the talkers

Figure A.2: Recording helmet prototypes.

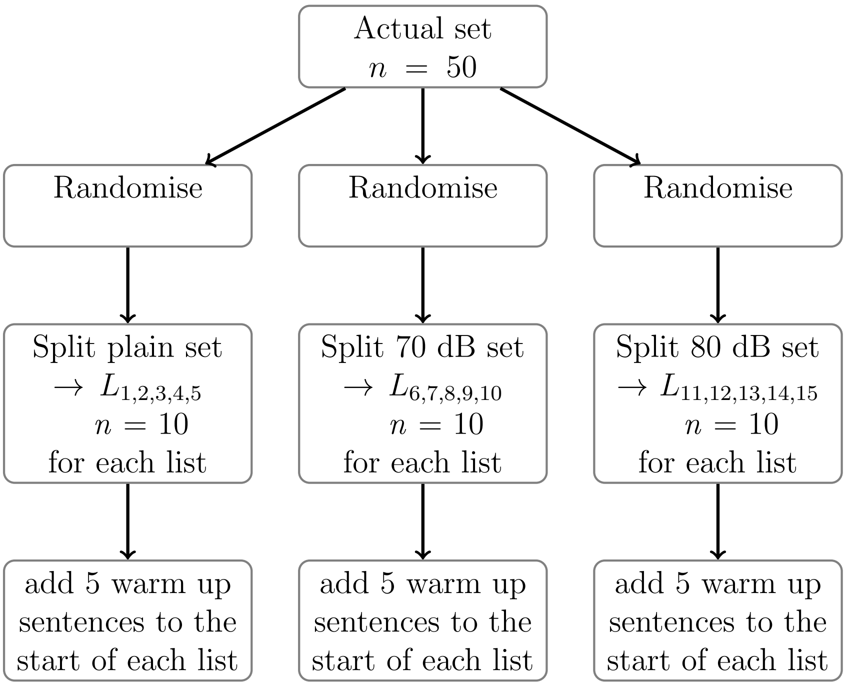


Figure A.3: Preparation of the prompt lists.

undergoing such prolonged recording would show vocal and auditory fatigue, especially in the later sessions;

- The effect of incorporating two masker levels, 80 and 70 dB SPL, in the recording procedure on the intensity of the articulatory modifications made by the talkers;

- Having a listener in the recording setup is important, because the Lombard effect is triggered as an unconscious reaction to noise, and by the need to maintain intelligible communication in noise [193]. A communication task was found to induce stronger audiovisual speech modifications than a reading task [102, 193]. Thus, the proposed communication protocol involved an interactive task in which the listener would listen to then respond to the recited sentences. Face-to-face is the communication modality chosen for this protocol. This is based on studies that reported increased visual speech saliency under such modality [91, 92]. The pilot study examined this communication protocol effect.

The results of this pilot study outline the dataset collection's procedure by selecting the number of recording sessions, the number of masker presentation levels, and the design of the communication protocol between the talker and the listener.

A.3.1 Method

Four male native speakers of British English from the staff and students at the University of Sheffield participated in this pilot study; three of them participated as talkers, each in the age range 18 – 23 years, and one 40-year-old participated as the listener. The subjects' hearing was screened using an on-line pure tone audiometric test [249]. Participants were paid for their contribution. Ethics permission for this study was obtained by following the University of Sheffield Ethics' Procedure.

Each talker took part in 15 recording sessions: five plain sessions with no masker presented; five sessions in which the masker was presented at 70 dB SPL, and another five sessions in which the masker was presented at 80 dB SPL. These sessions were organised in three blocks of recording (five sessions per block) to offer breaks for the talkers and the listener after each block. In each session, the talker uttered a prompt list of 15 sentences: five warm-up sentences followed by ten actual sentences. The order of the sessions, however, was different from one talker to another. Figure A.3 illustrated the creation of the prompt lists for the recording sessions: the actual sentence set designated to a talker was shuffled and decomposed into five prompt lists for the plain recording, then re-shuffled and decomposed into another five prompt lists for the Lombard recording at 70 dB, and re-shuffled again and decomposed into five prompt lists for the Lombard recording at 80 dB. Five sentences from the warm-up set were then added to each prompt list.

Figure 7.5 illustrates the collection setting. The audio and the video recording followed the same procedure as in Section 7.2.3. Each of the talkers was seated inside the booth and read sentences which appeared on the screen, while listening to the SSN masker. The listener was seated outside the booth facing a talker through a glass window. The listener maintained face-to-face contact while listening to the talker's speech presented at 60 dB SPL via a pair of headphones connected the audio box. Using a machine connected to the screen inside the booth, the listener provided feedback in which he entered the colour, the letter, and the digit in each sentence uttered by the talker using a labeled keyboard. Recording software controlled the presentation of the sentences at the talker's side and the collection of the feedback from the listener. At random points during the recording, the software prompted the listener to deliberately make errors while entering the keywords or ask the talker to repeat the sentences. The feedback on the error type (i.e., which keyword was misheard) or the repeat request by the listener appeared on the talker's screen. The talker responded to this alert by re-reading the sentence to the listener. The purpose of this step was to maintain the public Lombard loop which is driven by the communication need [176].

A.3.2 Results and Discussion

The vertical mouth aperture the talkers made during the speech production was used as a measure for the intensity of the visual Lombard speech. Faceware Analyser (FA) (Section 5.2) was used to extract the mouth landmarks from the videos that are necessary to compute the vertical mouth aperture for each frame (Landmark number 19 and 25 in the mouth region, Figure 5.1). The variance from the mean of the vertical mouth aperture made for each condition and at each session was then calculated.

Figure A.4 shows the vertical mouth aperture variance for each session. The results from Figure A.4 suggests a clear fatigue effect at the third block of the recordings for all talkers. Even for the 80 dB masker, the talkers showed a weaker Lombard response compared with the early blocks. For example, Talker 1 showed a weaker change in vertical mouth aperture in $block_3$- $session_{2,4}$ compared with $block_2$-$session_4$ and $block_1$- $session_4$. The number of errors made by the listener and the weight of these errors appears at the top of each Lombard session bar in Figure A.4. The weight of the error, as shown in Table A.1, is based on the error type. For example, when the listener reports an error in one keyword, the talker is informed of the error location in the uttered sentence (i.e., which Grid keyword the listener had misheard). The effort made by the talker when they re-read the misheard sentences is hypothesised to be directed to the location of the error, i.e., selective improvement to the sentence. On the other hand, when the listener makes a mistake in three keywords or requests the talker to repeat a sentence, the talker might try to make an overall improvement to the sentence intelligibility and hence induce more salient visual speech modification. Therefore, the intensity of the articulatory modification is a function of not only the masker SPL level, but also the number of errors that the listener made and the types of the errors. For example, Talker 1 in Figure A.4 showed stronger change in vertical mouth aperture for 70 dB SPL ($block_1$- $session_{2,5}$-error weight = (11, 17) , respectively) than under 80 dB SPL ($block_1$- $session_{1,4}$, error weight = (10, 11), respectively).

Figure A.5 shows the variance of the articulatory modification for each condition. Consistent with previous findings [164, 284, 301], talkers showed salient visual speech modification at 80 dB SPL, while they responded differently to the 70 dB SPL masker. For example, Talker 1 showed a nearly linear response to the increase in the SPL level; Talker 2 showed a comparable effect under 70 and 80 dB; and Talker 3 showed very

Error type	one keywords	Two keywords	Three keywords	Repeat request
Weight	1	2	3	3
Number of error	a	b	c	d
Error weight	a	2b	3c	3d
Session x errors	([a+b+c+d], [a+2b+3c+3d])			

Table A.1: Calculating the error weight based on the error type in session x.

poor modification response under 70 dB. Such results are driven by the variability in Lombard effect response between talkers [152].

Implications on the Collection Design Given the results of this pilot study, a number of points will be reflected in the dataset collection procedure's design:

- Given the results in Figure A.4, the recording duration was shortened from three blocks to two blocks to control the effect of the auditory and vocal fatigue on the talkers' production.
- Given the variations in visual speech modification made by the talkers in response to the masker presented at 70 dB SPL (Figure A.4), the number of pressure levels presented in the experiment was reduced into one level, that is 80 dB SPL.
- The communication protocol needs to be revisited and modified; the talkers reported that seeing the listener informing errors have induced some negative feelings such as frustration and embarrassment. To control the psychological effects that may result from seeing the listener, a white screen was placed on the booth's window to limit face-to-face interaction. To control speech modification levels, talkers should refrain from knowing the error type, as it might have a direct effect on the hyper-articulation energy. The number of the errors made by the listener should be also uniform across all Lombard sessions.

A.4 Phoneme Viewer

An interactive MATLAB software named Phoneme Viewer, was developed to visualise the articulatory features of phonemes extracted from the utterances pool. The class diagram in Figure A.6 illustrates the structure of the software. The software utilises the phonetic alignment of the utterances to extract and label phonemes frames. To

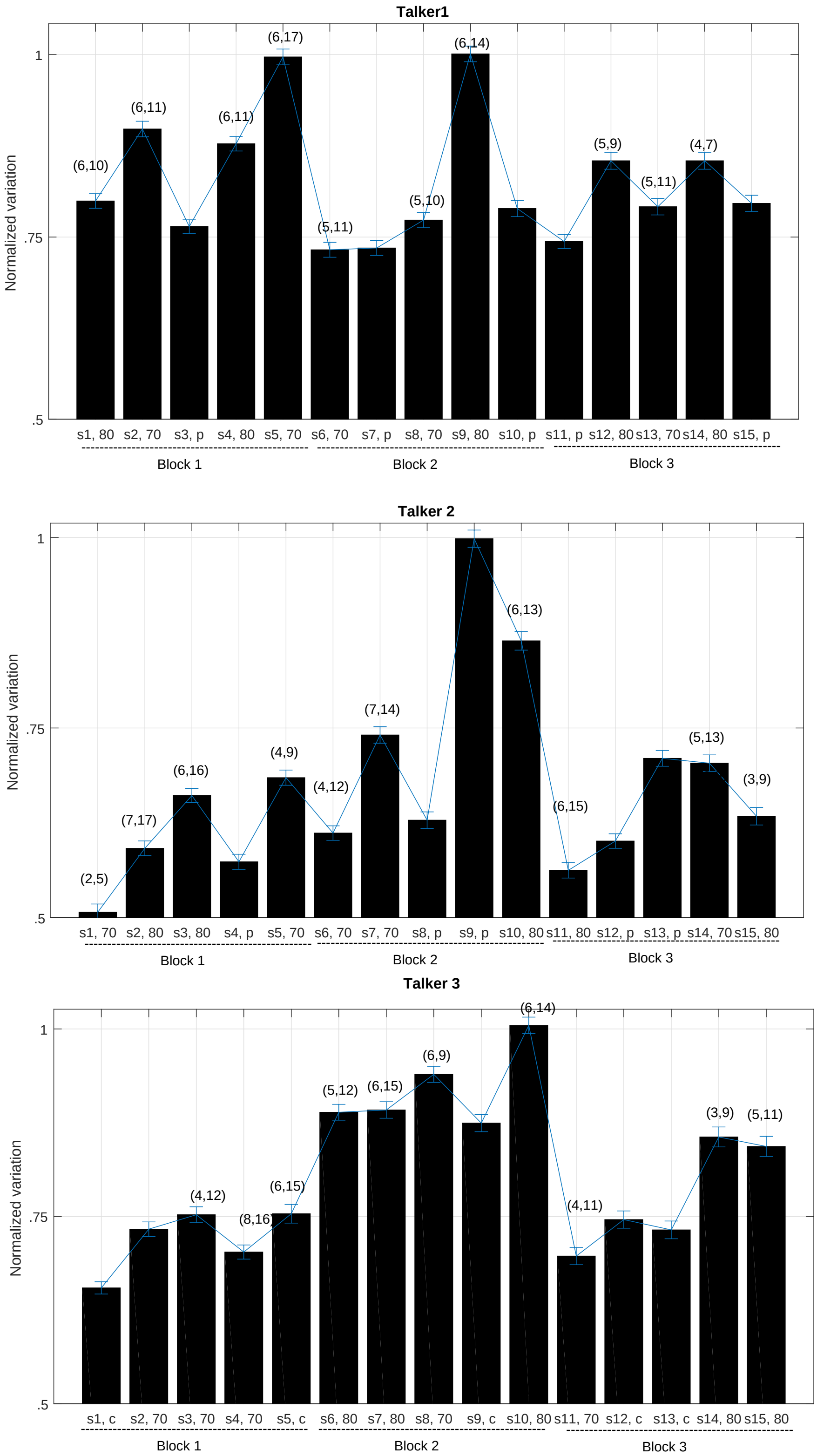


Figure A.4: The effect of the recording duration and the communication task on the variation of vertical mouth aperture. (number of Errors, Errors weight) by the listener at each session is displayed on each session bar.

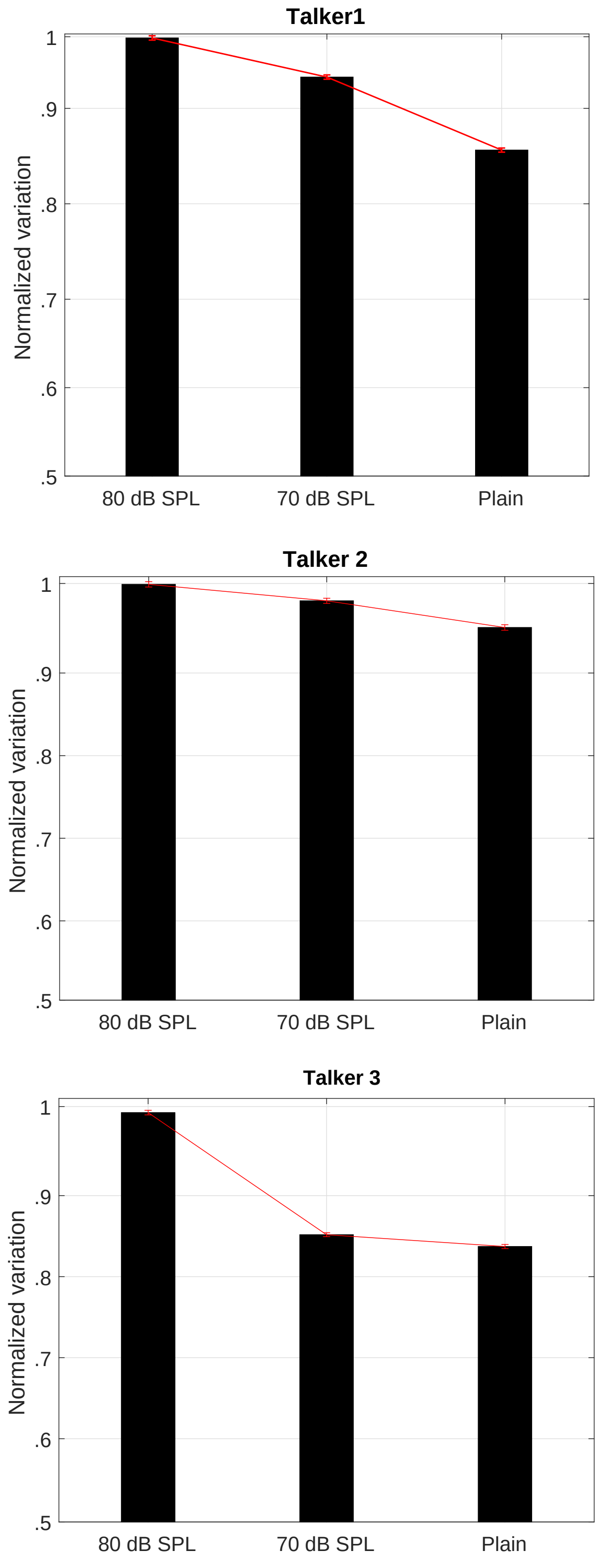


Figure A.5: The effect of the masker pressure level on the variation of vertical mouth aperture.

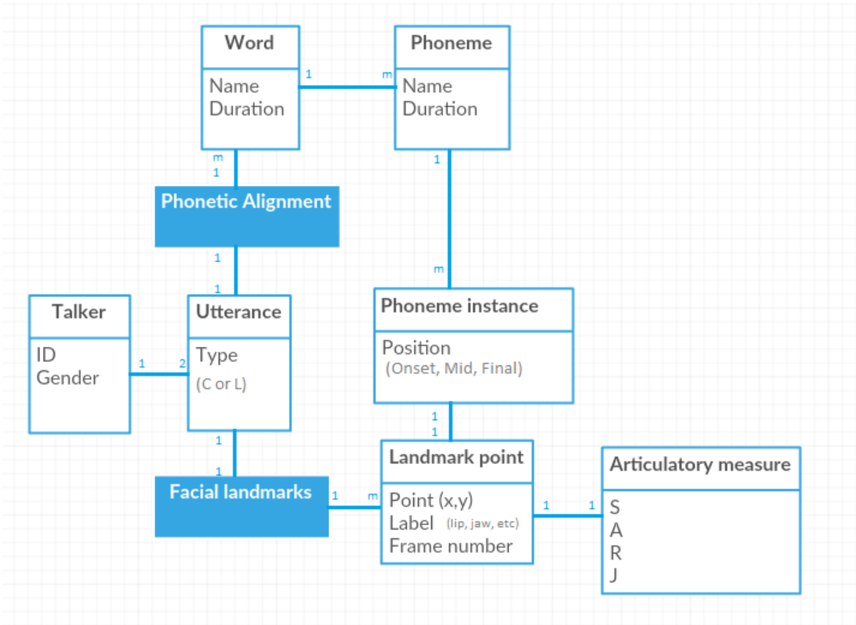


Figure A.6: Phoneme Viewer class diagram.

elaborate on the labeling process, consider a phoneme X that occurs n times in an utterances group; in each occurrence, phoneme X is expressed by a stream of frames that illustrates the time-line of phoneme X's production. The tool labels each phoneme frame according to it's position in the corresponding stream (i.e., as an onset frame in the production, as a middle frame, or as the final frame in the stream).

The software provides two cases to visualise the phonemes. The first case (Figure A.7- top) is concerned with visualising phonemes at the utterance level across all talkers irrespective of the context in which they were presented in. The second case (Figure A.7- bottom) looks into the behaviour of phonemes in a similar context, a word in this case, between the talkers.

In the utterance level view, the software enables four different views of the data from all occurrences of a selected phoneme:

- 'All': considers all frames;
- 'Mean mid': considers the average articulatory features of the three mid frames;
- 'Mid': considers all mid frames;
- 'Onset': considers only the onset frames;
- 'Final': considers only the final frames.

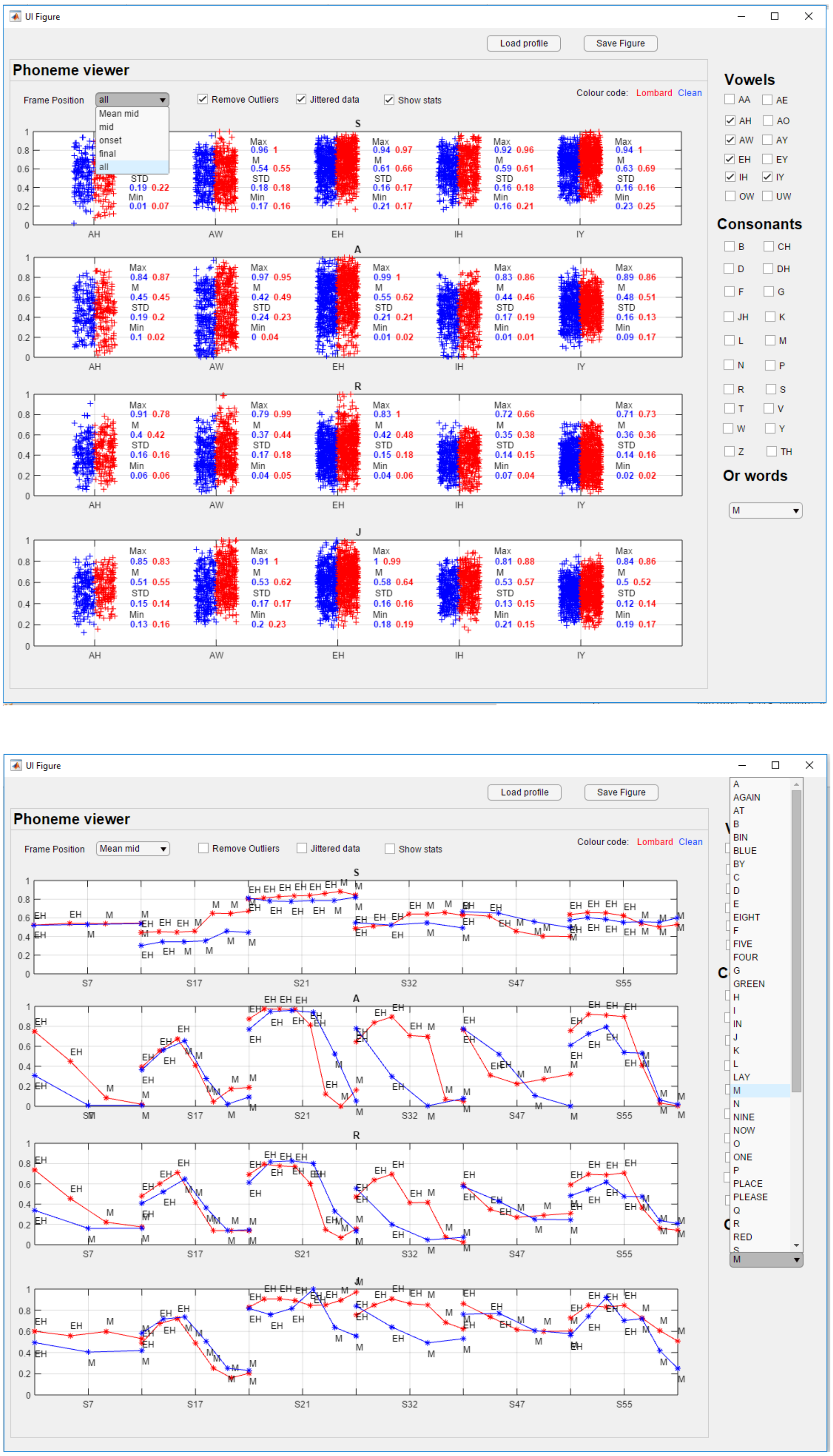


Figure A.7: Visual analytic app interface.

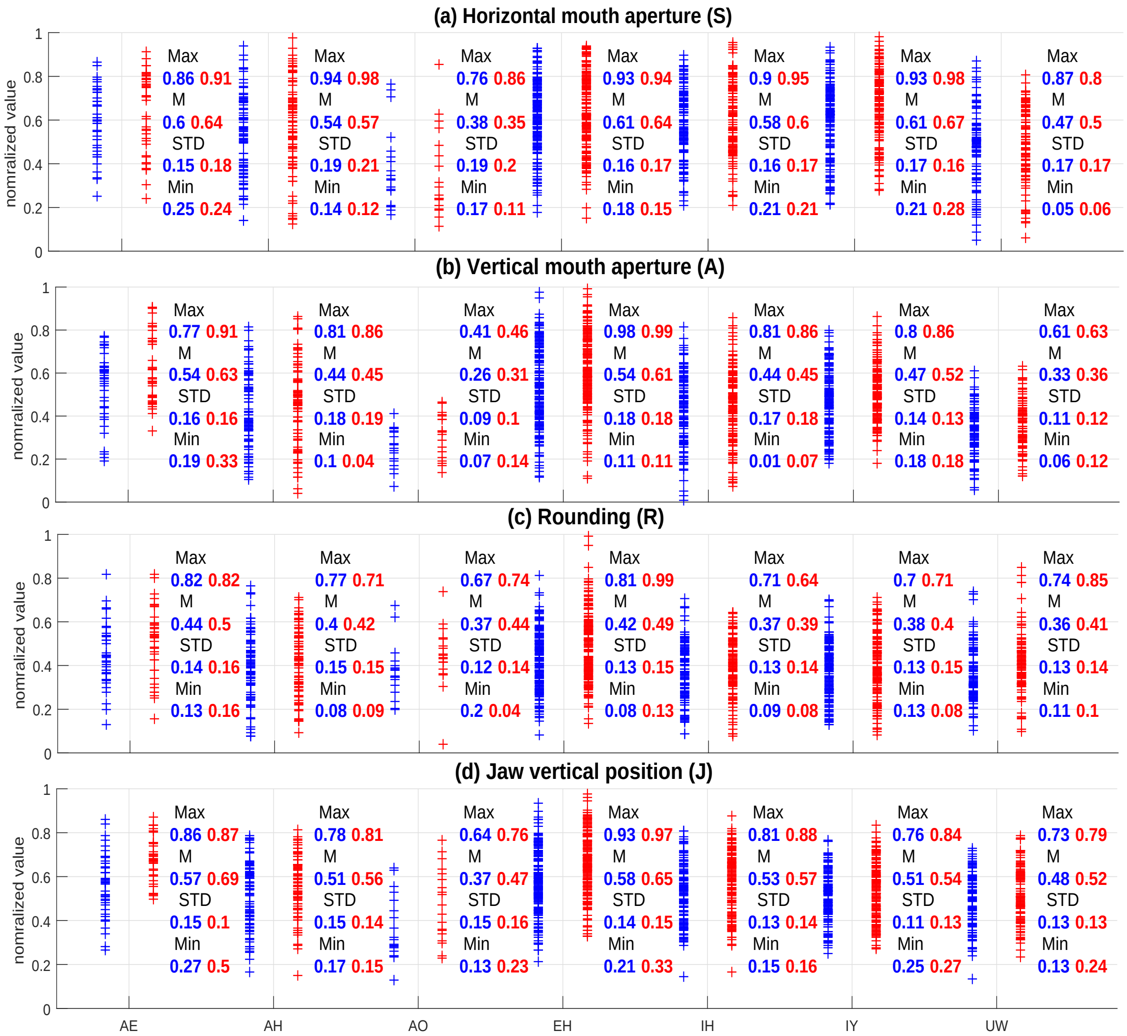


Figure A.8: Monophthong vowels. Blue: plain, red: Lombard

The desired phonemes can be selected from the right pane of software. Upon selection, the four articulatory measures (explained in Chapter 7) of the selected phonemes are displayed in the four centre plots in plain and in Lombard conditions. Two different visualisation methods are provided: a line plot and a jitter plot. The software also enables the data extremes (outliers) to be removed using the Interquartile Range Rule by setting minimum and maximum values for the presented data using the first and third quartiles, and then removing all extremes less than the minimum and greater than the maximum. A statistics summary that includes the mean, the standard deviation, and the minimum and maximum values of the plotted phonemes data can be also added to the plot area.

In the second visualisation case, i.e., the word level view, common words that

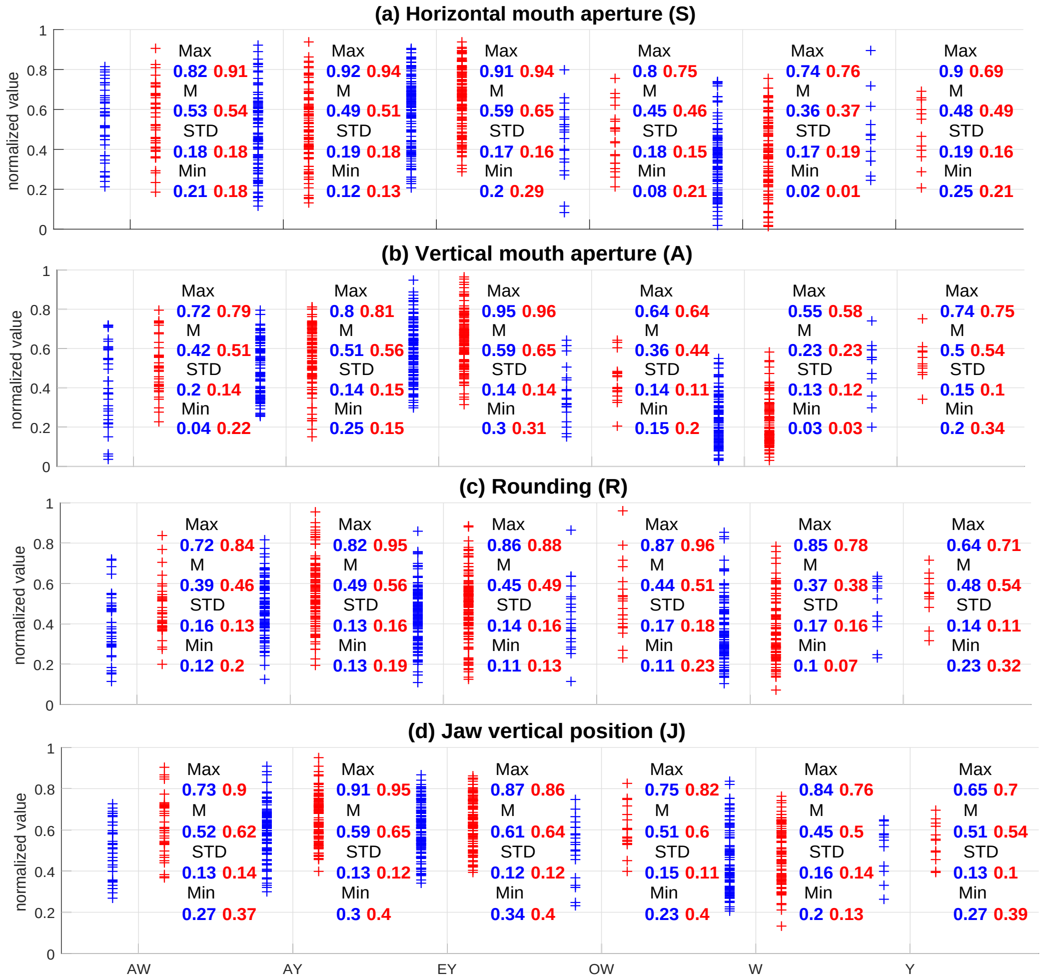


Figure A.9: Diphthong and semi-vowels. Blue: plain, red: Lombard.

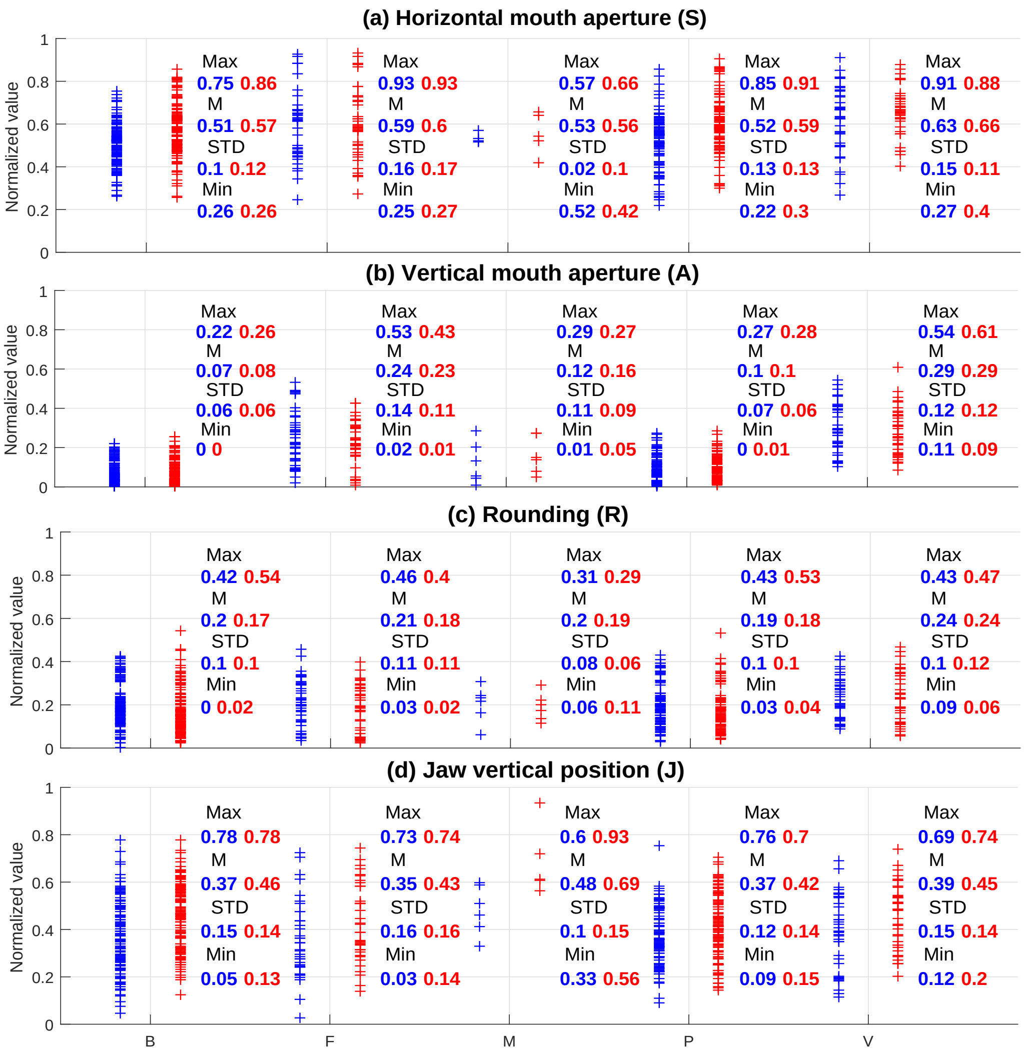


Figure A.10: Labials: bilabial and labiodental consonants. Blue: plain, red: Lombard.

Figure A.11: Coronal: dental, alveolar and palato-alveolar consonants. Blue: plain, red: Lombard.

have been articulated by all takers can be selected from a drop-down list. Upon selection, the articulatory measures taken from all phoneme frames from the selected word are plotted in plain and Lombard conditions against their talkers. This view enables the comparison of articulation behaviour across talkers. Table A.2 shows the sentence list from which the letters were extracted from each talker. This illustrates each letter's neighbors.

The software uses the Arpabet phoneme notation [168]. Table 7.2 maps between the Arpabet notation and the IPA notation. The software saves all plots to figures that can easily be imported to the Latex environment. In the following section, figures featuring utterance level and word level visualisation are presented.

Letter	S7	S14
A	bin red in A 7 now	lay green by A 4 now
E	lay red in E 5 soon	set green with E 5 now
I	bin green by I 4 soon	bin white by I 2 again
O	bin white at O 2 now	bin white with O 2 soon
B	lay green by B 4 again	bin white with B 5 now
N	place green at N 3 please	lay white with N 9 again
T	place white with T 2 again	bin white in T 2 please
C	place blue in C 0 soon	set blue at C 1 now
Letter	S21	S32
A	bin red in A 9 please	lay red by A 7 please
E	set white with E 4 now	lay blue at E 7 again
I	bin green at I 6 now	set white at I 8 now
O	set blue with O 8 please	bin green at O 3 please
B	lay green by B 4 now	bin green in B 6 now
N	place green at N 3 again	place white with N 1 now
T	place green at T 5 now	lay red in T 7 please
C	place blue at C 7 soon	set blue in C 6 please
Letter	S44	S46
A	bin blue at A 5 soon	lay white with A 5 please
E	place green in E 6 soon	place red at E 6 please
I	bin red at I 2 please	bin white at I zero now
O	set blue in O 6 please	set blue by O 1 now
B	bin blue by B 3 again	set white at B 9 soon
N	bin blue in N 4 please	lay white with N 2 soon
T	lay green in T 4 soon	set blue by T 2 soon
C	set red in C 7 soon	set red with C 4 again
Letter	S47	55
A	lay red in A 1 please	bin red at A 8 please
E	lay red in E 6 now	place white at E 2 again
I	set red with I 7 soon	set red in I zero please
O	bin white in O 7 now	set red in O one again
B	bin blue at B 6 please	lay green with B 7 please
N	bin blue at N 7 please	place blue in N 5 now
T	lay white by T 5 soon	lay white in T 4 now
C	bin red by C 1 again	lay white with C 1 again

Table A.2: Sentence lists of the selected letters.

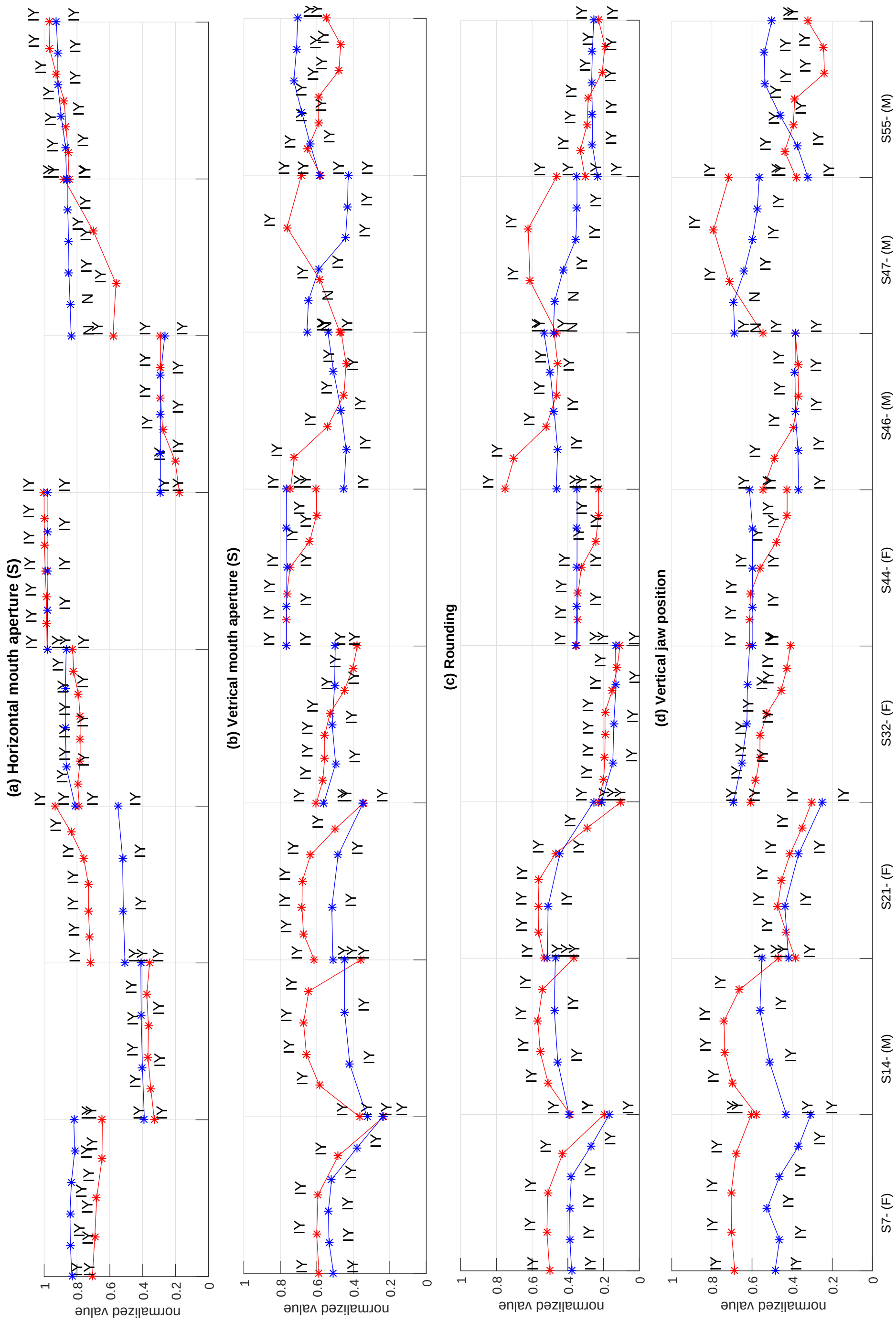


Figure A.12: Letter E. Blue: plain, red: Lombard.

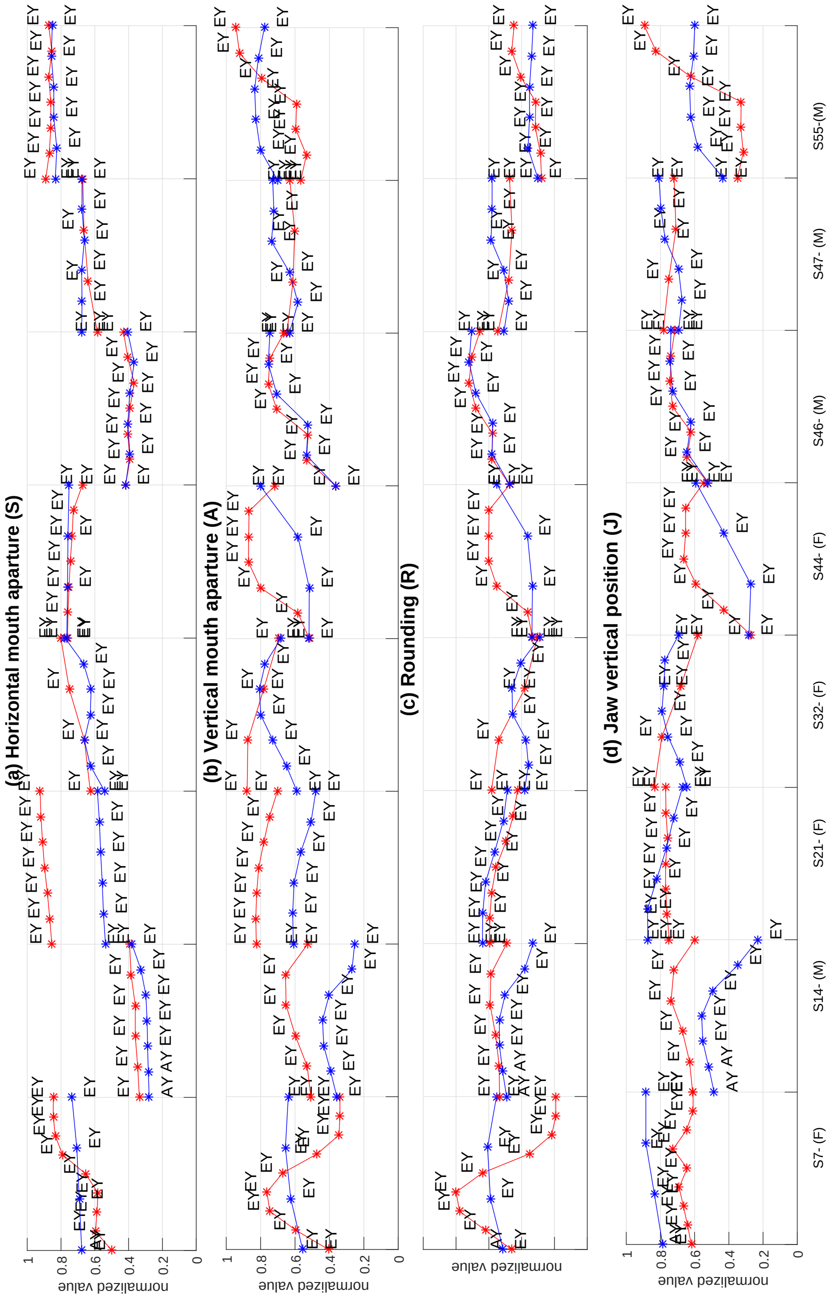


Figure A.13: Letter A. Blue: plain, red: Lombard.

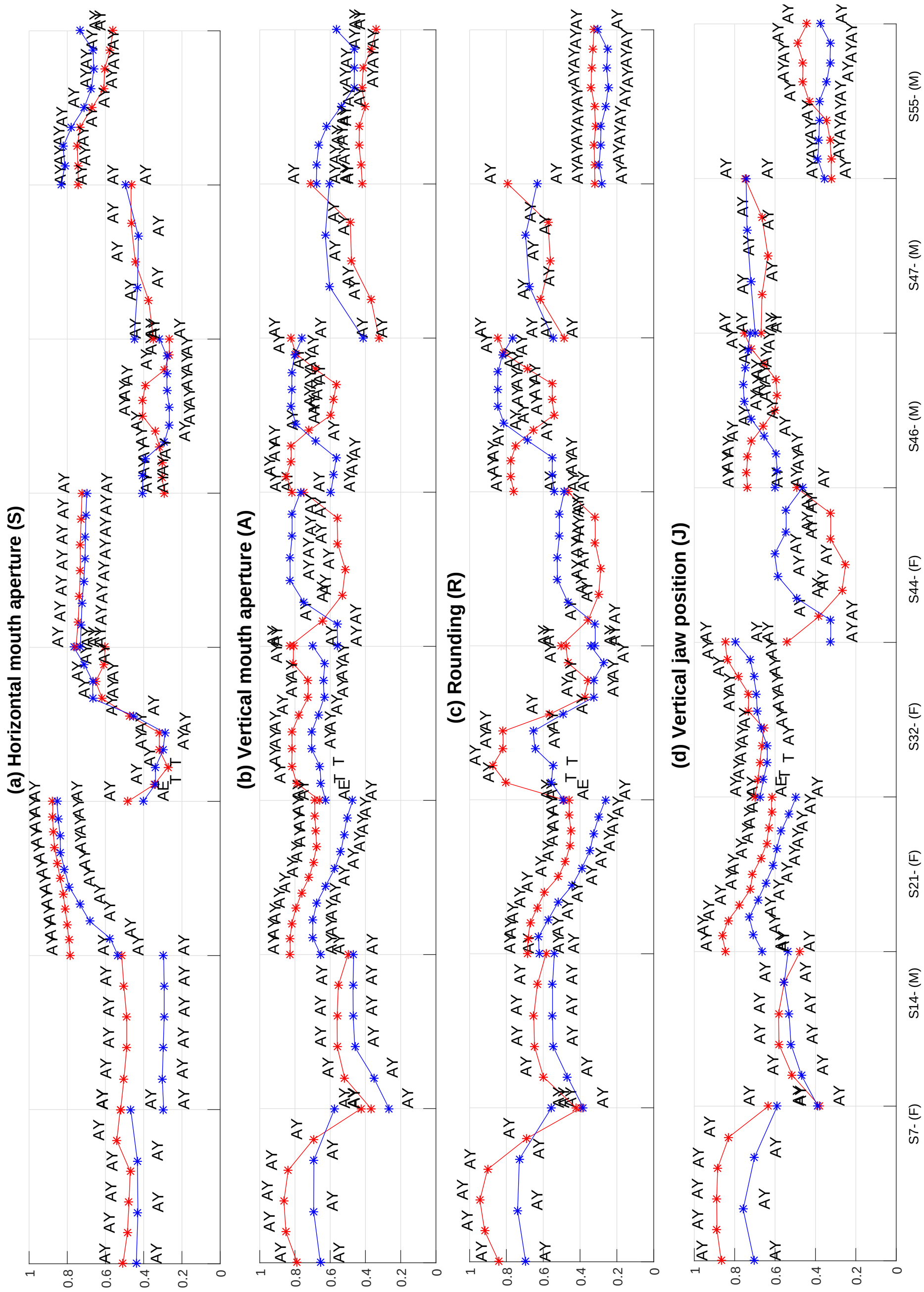


Figure A.14: Letter I. Blue: plain, red: Lombard.

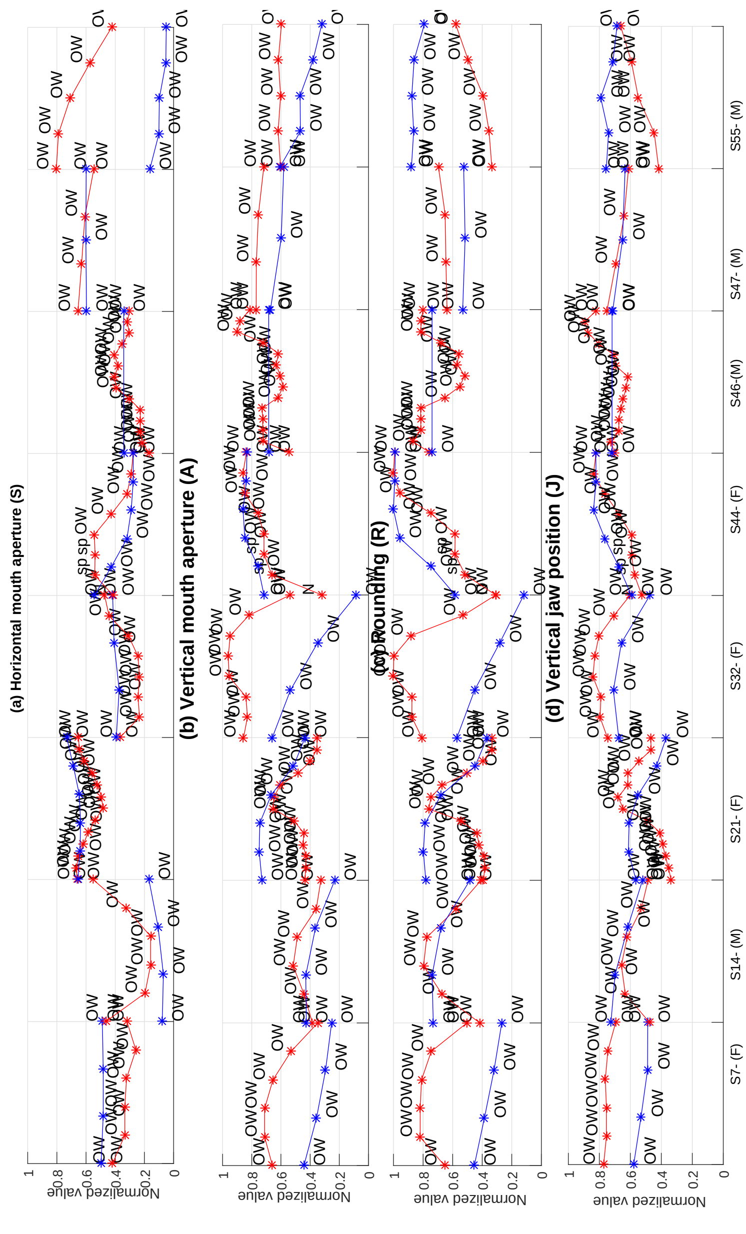


Figure A.15: Letter O. Blue: plain, red: Lombard.

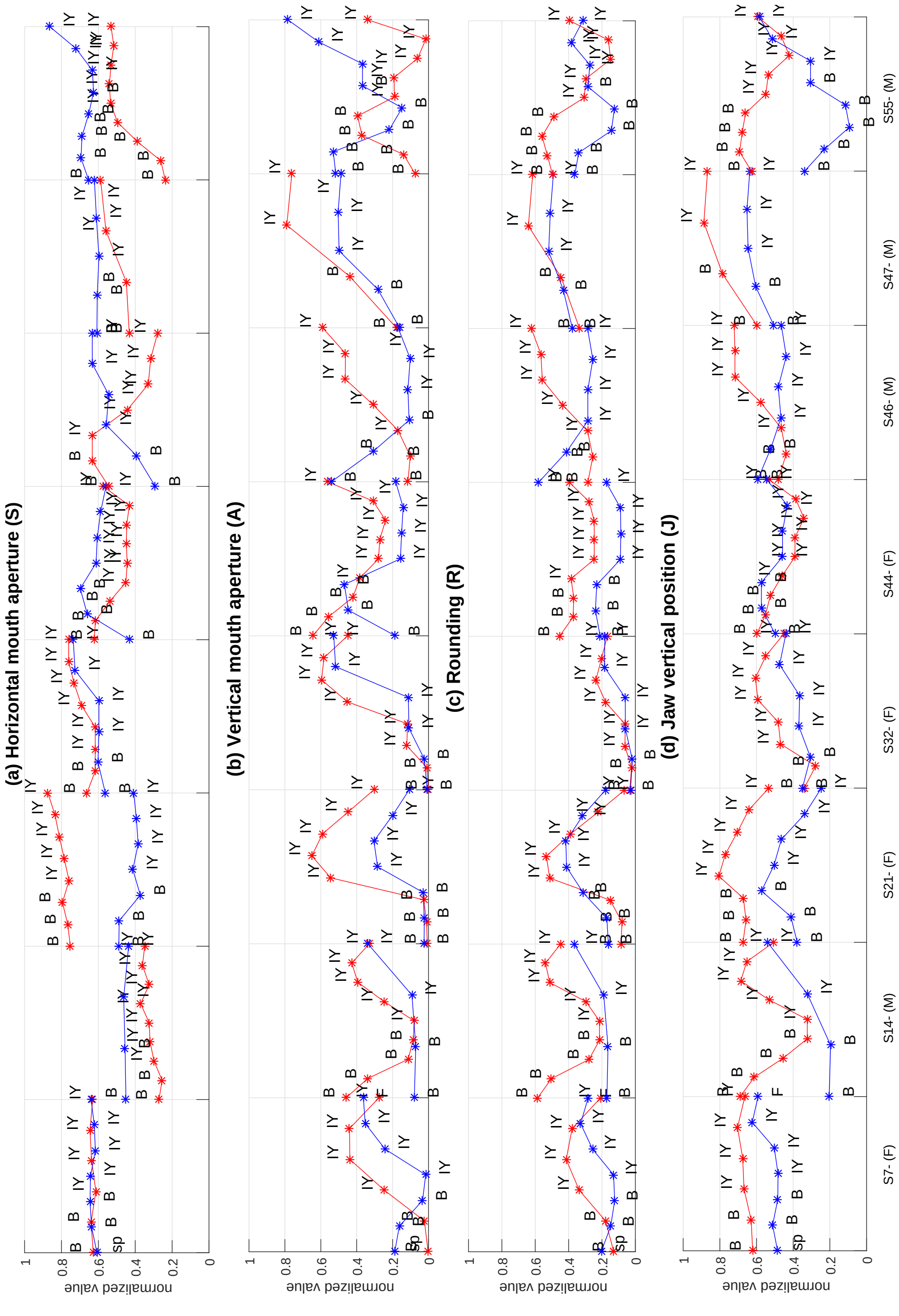


Figure A.16: Letter B. Blue: plain, red: Lombard.

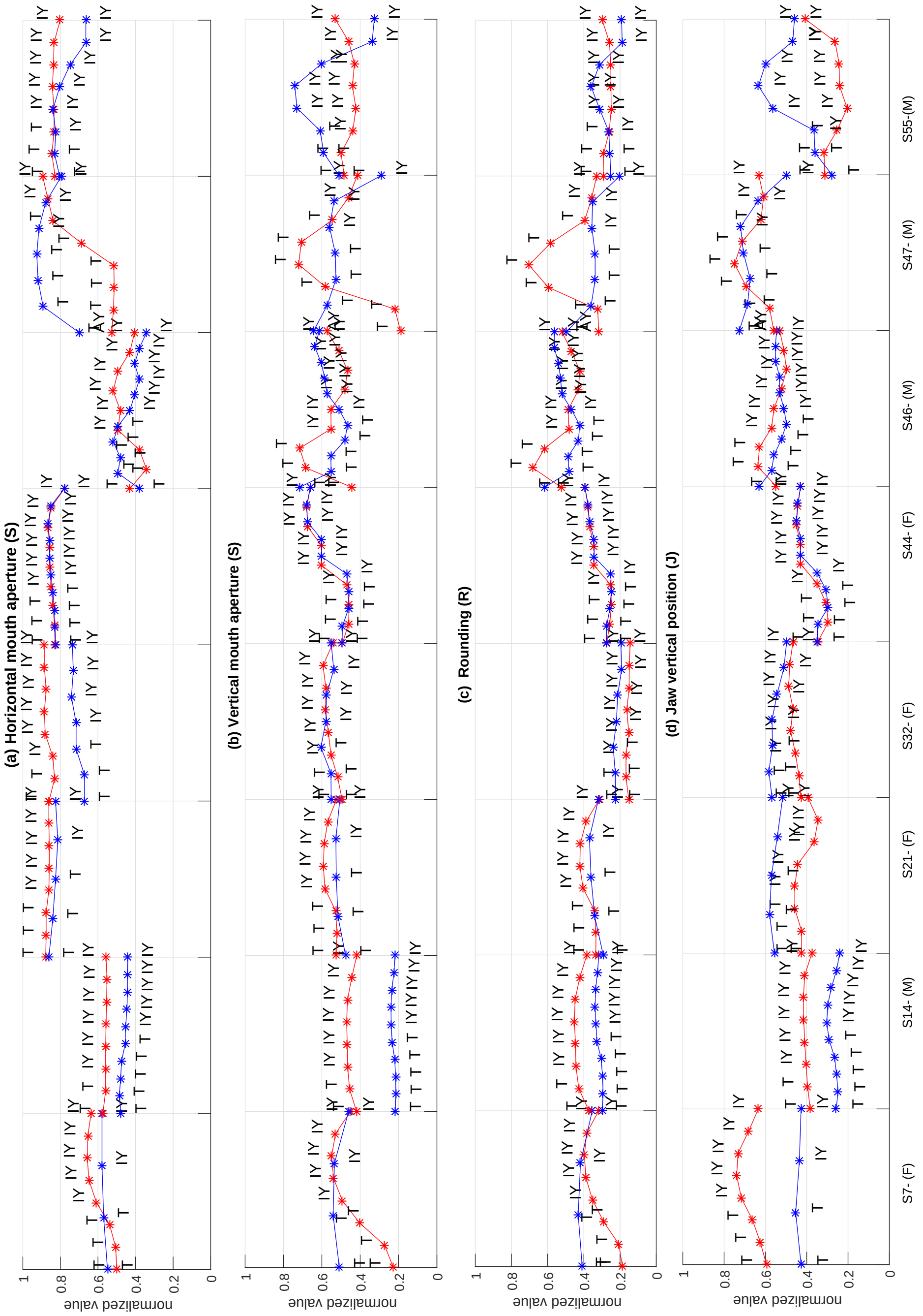


Figure A.17: Letter T. Blue: plain, red: Lombard.

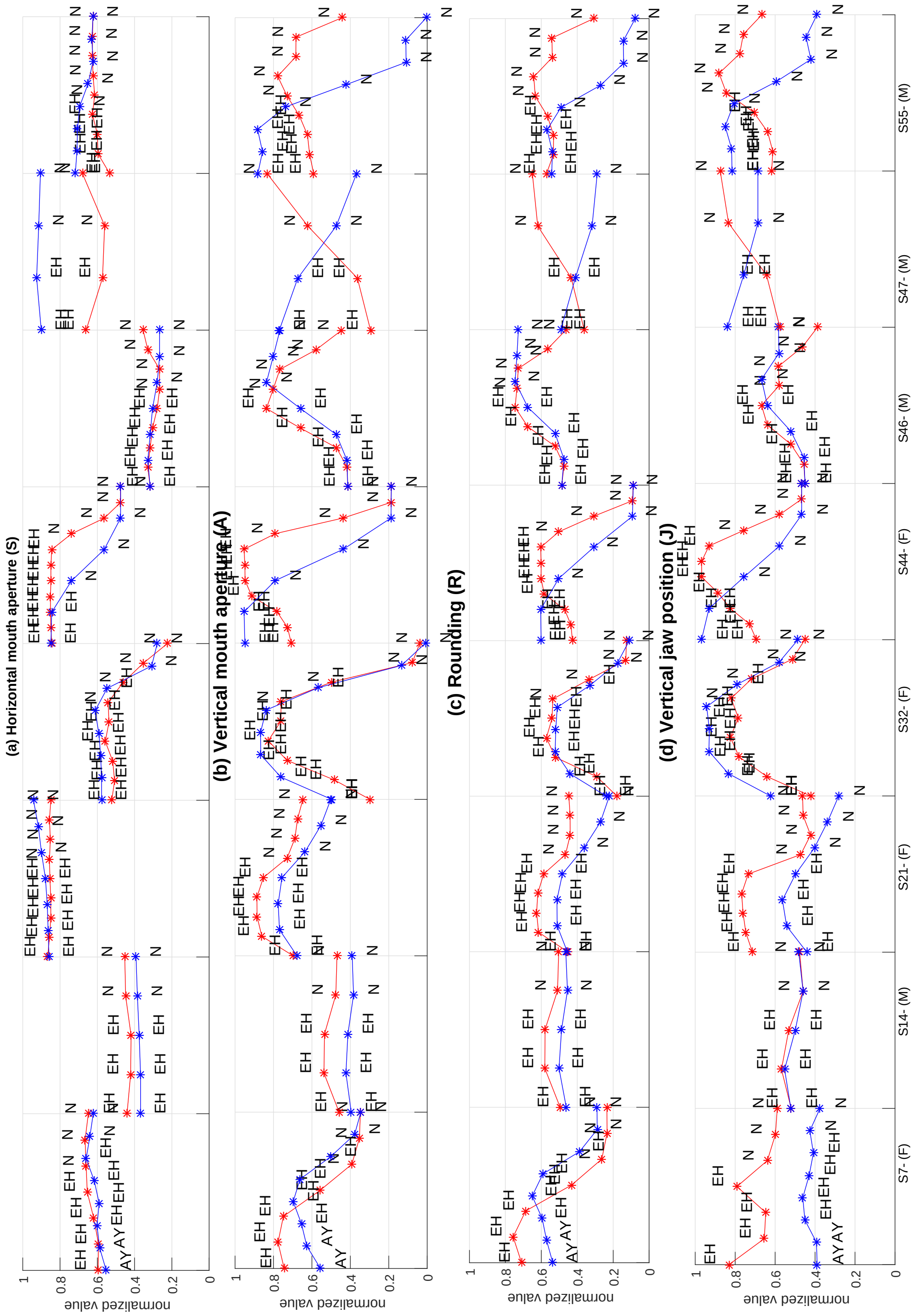


Figure A.18: Letter N. Blue: plain, red: Lombard.

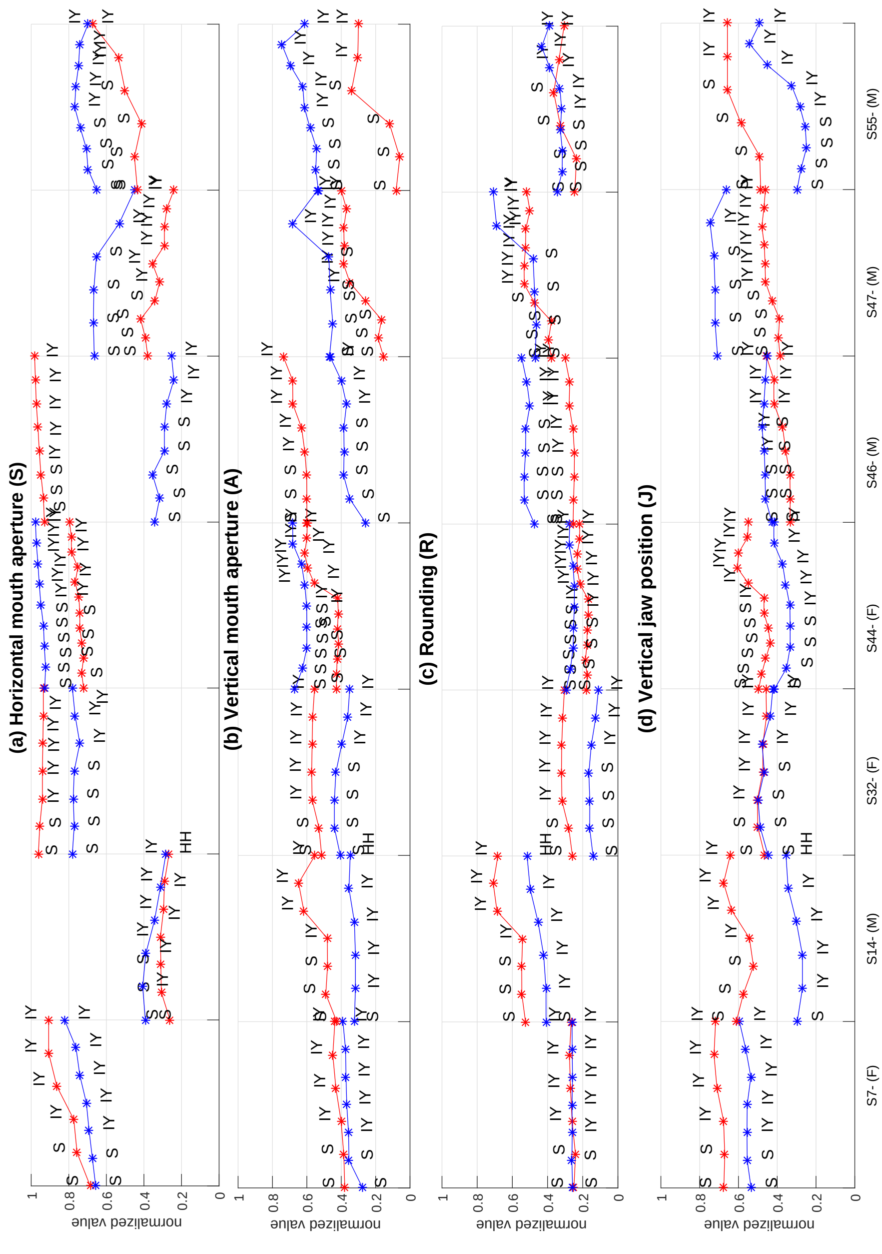


Figure A.19: Letter C. Blue: plain, red: Lombard.

Bibliography

[1] Avicar project: Audio-visual speech recognition in a car. http://www.isle.illinois.edu/sst/AVICAR. Accessed: 2014-09-16.

[2] Faceware analyser, professional facial tracking software. http://facewaretech.com/products/software/analyzer. Accessed: 2017-09-24.

[3] Intel realsense sdk. https://software.intel.com/en-us/intel-realsense-sdk. Accessed: 2017-09-24.

[4] Visage technologies: Face tracking and analysis. http://visagetechnologies.com/products-and-services/visagesdk. Accessed: 2017-09-24.

[5] Andrew Abel and Amir Hussain. *Audio and Visual Speech Relationship Cognitively Inspired Audiovisual Speech Filtering*. Springer, 5–12, 2015.

[6] Andrew Abel and Amir Hussain. *Cognitively Inspired Audiovisual Speech Filtering: Towards an Intelligent, Fuzzy Based, Multimodal, Two-Stage Speech Enhancement System*. Springer, 1st edition, 2015.

[7] Ahsan Adeel, Mandar Gogate, and Amir Hussain. Towards next-generation lip-reading driven hearing-aids: A preliminary prototype demo. In *1st International Workshop on Challenges in Hearing Assistive Technology (CHAT-2017)*, 2017.

[8] Ali Adjoudani and Christian Benoit. On the integration of auditory and visual parameters in an hmm-based asr. In *Speechreading by humans and machines*, pages 461–471. Springer, 1996.

[9] Merav Ahissar and Shaul Hochstein. The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, 8(10):457–464, 2004.

[10] Najwa Al-Ghamdi, Abeer Al-Nafjan, and Yousef Al-Ohali. A computer-based aural rehabilitation assistant for cochlear implanted children. In *Proceeding of International Conference on Future Information Technology*, volume 13, pages 279–284, 2011.

[11] Claude Alain, Joel S Snyder, Yu He, and Karen S Reinke. Changes in auditory cortex parallel rapid perceptual learning. *Cerebral Cortex*, 17(5):1074–1084, 2007.

[12] M Alex Meredith and Barry E Stein. Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain research*, 365(2):350–354, 1986.

[13] Simon Alexanderson and Jonas Beskow. Animated lombard speech: motion capture, facial animation and visual intelligibility of speech produced in adverse conditions. *Computer Speech and Language*, 28(2):607–618, 2014.

[14] Rabab Alghady, Yoshihiko Gotoh, and Steve Maddock. Analysis of visemes in the grid corpus. In *UKSpeech 2016 Conference at the University of Sheffield*.

[15] Najwa Alghamdi, Steve Maddock, Guy J. Brown, and Jon Barker. A comparison of audiovisual and auditory-only training on the perception of spectrally-distorted speech. In *The International Congress of Phonetic Sciences (ICPhS)*, 2015.

[16] Ibrahim Almajai and Ben Milner. Enhancing audio speech using visual speech features. In *INTERSPEECH 2009: Tenth Annual Conference of the International Speech Communication Association*, 2009.

[17] Deborah K Amazi and Sharon R Garber. The lombard sign as a function of age and task. *Journal of Speech, Language, and Hearing Research*, 25(4):581–585, 1982.

[18] Carly A Anderson, Ian M Wiggins, Pádraig T Kitterick, and Douglas EH Hartley. Adaptive benefit of cross-modal plasticity following cochlear implantation in deaf adults. *Proceedings of the National Academy of Sciences*, pages 10256–10261, 2017.

[19] Peter Assmann and Quentin Summerfield. The perception of speech under adverse conditions. In *Speech processing in the auditory system*, pages 231–308. Springer, 2004.

[20] Mercedes Atienza, Jose L Cantero, and Elena Dominguez-Marin. The time course of neural changes underlying auditory perceptual learning. *Learning and Memory*, 9(3):138–150, 2002.

[21] Karen Banai and Sygal Amitay. Stimulus uncertainty in auditory perceptual learning. *Vision research*, 61:83–88, 2012.

[22] Fabrice Bellard, M Niedermayer, et al. Ffmpeg. *Availabel from: http://ffmpeg.org*, 2012.

[23] John Bench, Åse Kowal, and John Bamford. The bkb (bamford-kowal-bench) sentence lists for partially-hearing children. *British journal of audiology*, 13(3):108–112, 1979.

[24] Tessa Bent, Adam Buchwald, and David B Pisoni. Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *The Journal of the Acoustical Society of America*, 126(5):2660–2669, 2009.

[25] Kenneth Walter Berger. *Speechreading: Principles and methods.* National Educational Press, 1972.

[26] Robert I Bermant and Robert B Welch. Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition. *Perceptual and Motor Skills*, 43(2):487–493, 1976.

[27] Joshua GW Bernstein and Ken W Grant. Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 125(5):3358–3372, 2009.

[28] Lynne E Bernstein, Edward T Auer Jr, Silvio P Eberhardt, and Jintao Jiang. Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Frontiers in neuroscience*, 7, 2013.

[29] Lynne E. Bernstein and Einat Liebenthal. Neural pathways for visual speech perception. *Frontiers in Neuroscience*, 8:01–18, 2014.

[30] Paul Bertelson, Jean Vroomen, and Béatrice De Gelder. Visual recalibration of auditory speech identification a mcgurk aftereffect. *Psychological Science*, 14(6):592–597, 2003.

[31] Turki A Binturki. *Analysis of pronunciation errors of Saudi ESL learners*. PhD thesis, Southern Illinois University at Carbondale, 2008.

[32] P Boersma and D Weenink. Praat speech processing software. *Institute of Phonetics Sciences of the University of Amsterdam. http://www. praat. org*, 2001.

[33] Daniel J Bosnyak, Robert A Eaton, and Larry E Roberts. Distributed auditory cortical representations are modified when non-musicians are trained at pitch discrimination with 40 hz amplitude modulated tones. *Cerebral Cortex*, 14(10):1088–1099, 2004.

[34] Ann R Bradlow, David B Pisoni, Reiko Akahane-Yamada, and Yoh'ichi Tohkura. Training japanese listeners to identify english/r/and/l: Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, 101(4):2299–2310, 1997.

[35] Ann R Bradlow, Gina M Torretta, and David B Pisoni. Intelligibility of normal speech: Global and fine-grained acoustic-phonetic talker characteristics. *Speech communication*, 20(3):255–272, 1996.

[36] M Breeuwer and Reinier Plomp. Speechreading supplemented with frequency-selective sound-pressure information. *The Journal of the Acoustical Society of America*, 76(3):686–691, 1984.

[37] Luke J Brook. *Analysing a new mobile bilateral audiology test for children*. PhD thesis, SRI Security Research Institute, Edith Cowan University, Perth, Western Australia, 2013.

[38] Luke J Brook and Patricia AH Williams. Developing a mobile audiometric sound booth application for apple ios devices. In *The 2nd Australian eHealth Informatics and Security Conference*, 2013.

[39] NM Brooke and Quentin Summerfield. Analysis, synthesis, and perception of visible articulatory movements. *Journal of phonetics*, 11(1):63–76, 1983.

[40] Armin Bruderlin and Lance Williams. Motion signal processing. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 97–104. ACM, 1995.

[41] Henrik Brumm and Sue Anne Zollinger. The evolution of the lombard effect: 100 years of psychoacoustic research. *Behaviour*, 148(11-13):1173–1198, 2011.

[42] Rachel Caissie. Auditory training: From speech sounds to heart sounds. In *Teaching Heart Auscultation to Health Professionals*, chapter 5. Dalhousie University Press, Dalhousie University, 2011.

[43] Emanuela Magno Caldognetto, Giulio Perin, and Claudio Zmarich. Labial coarticulation modeling for realistic facial animation. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, pages 505–511. IEEE Computer Society, 2002.

[44] MB Calford. Dynamic representational plasticity in sensory cortex. *Neuroscience*, 111(4):709–738, 2002.

[45] Gemma Calvert and Ruth Campbell. Reading speech from still and moving faces: the neural substrates of visible speech. *Cognitive Neuroscience, Journal of*, 15(1):57–70, 2003.

[46] Gemma A Calvert, Edward T Bullmore, Michael J Brammer, Ruth Campbell, Steven CR Williams, Philip K McGuire, Peter WR Woodruff, Susan D Iversen, and Anthony S David. Activation of auditory cortex during silent lipreading. *science*, 276(5312):593–596, 1997.

[47] R Campbell and T-JE Mohammed. Speechreading for information gathering: a survey of scientific sources. Deafness Cognition and Language (DCAL) Research Centre, Division of Psychology and Language Sciences, University College London, 2010.

[48] Ruth Campbell. The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1493):1001–1010, 2008.

[49] David Carmel and Marisa Carrasco. Perceptual learning and dynamic changes in primary visual cortex. *Neuron*, 57(6):799–801, 2008.

[50] Patrick Cavanagh and Yvan G Leclerc. Shape from shadows. *Journal of Experimental Psychology: Human Perception and Performance*, 15(1):3, 1989.

[51] Baoquan Chen, Frank Dachille, and Arie Kaufman. Forward image mapping. In *Proceedings of the conference on Visualization'99: celebrating ten years*, pages 89–96. IEEE Computer Society Press, 1999.

[52] Charles TM Choi and Yi-Hsuan Lee. A review of stimulating strategies for cochlear implants. In *Cochlear Implant Research Updates*. InTech, 2012.

[53] Michael M Cohen, Dominic W Massaro, et al. Modeling coarticulation in synthetic visual speech. *Models and techniques in computer animation*, 92:139–156, 1993.

[54] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[55] Martin Cooke, Simon King, Maëva Garnier, and Vincent Aubanel. The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech and Language*, 28(2):543–571, 2014.

[56] Martin Cooke, ML Garcia Lecumberri, and Jon Barker. 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, 2008.

[57] Tim Cootes. An introduction to active shape models. *Image Processing and Analysis*, pages 223–248, 2000.

[58] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[59] David Cornish and Diane Dukette. *The Essential 20: Twenty Components of an Excellent Health Care Team*. Rosedog Press, 2009.

[60] Nicola Daly, John Bench, and Hilary Chappell. Gender differences in visual speech variables. *Journal-Academy of Rehabilitative Audiology*, 30:63–76, 1997.

[61] Jess Dancer, Mark Krain, Carolyn Thompson, and Priscilla Davis. A cross-sectional investigation of speechreading in adults: effects of age, gender, practice, and education. *The Volta Review*, 96(1):31–40, 1994.

[62] Chris Davis and Jeesun Kim. Is speech produced in noise more distinct and/or consistent? *Speech Science and Technology*, pages 46–49, 2012.

[63] Chris Davis, Jeesun Kim, Katja Grauwinkel, and Hansjörg Mixdorff. Lombard speech: Auditory (a), visual (v) and av effects. In *In Proceedings of the Third International Conference on Speech Prosody*, pages 248–252. TUD Press, 2006.

[64] Chris Davis, Amanda Sironic, and Jeesun Kim. Perceptual processing of audiovisual lombard speech. In *Proceedings of the 11th Australian International Conference on Speech Science and Technology*, 2006.

[65] Matthew H Davis, Ingrid S Johnsrude, Alexis Hervais-Adelman, Karen Taylor, and Carolyn McGettigan. Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, 134(2):222, 2005.

[66] Beatrice de Gelder, Paul Bertelson, Jean Vroomen, and Hsuan Chin Chen. Inter-language differences in the mcgurk effect for dutch and cantonese listeners. In *the European Conference on Speech Communication and Technology (EUROSPEECH)*, 1995.

[67] Beatrice De Gelder and Jean Vroomen. The perception of emotions by ear and by eye. *Cognition and Emotion*, 14(3):289–311, 2000.

[68] Salil Prashant Deena. *Visual speech synthesis by learning joint probabilistic models of audio and video.* PhD thesis, The University of Manchester, 2012.

[69] Karine Delhommeau, Christophe Micheyl, and Roland Jouvent. Generalization of frequency discrimination learning across frequencies and ears: implications for underlying neural mechanisms in humans. *Journal of the Association for Research in Otolaryngology*, 6(2):171–179, 2005.

[70] Laurent Demany. Perceptual learning in frequency discrimination. *The Journal of the Acoustical Society of America*, 78(3):1118–1120, 1985.

[71] Laurent Demany and Catherine Semal. Learning to perceive pitch differences. *The Journal of the Acoustical Society of America*, 111(3):1377–1388, 2002.

[72] Peter B Denes and Elliot Pinson. *The speech chain.* Macmillan, 1993.

[73] Zhigang Deng and Junyong Noh. Computer facial animation: A survey. In *Data-driven 3D facial animation*, pages 1–28. Springer, 2008.

[74] Sheetal Desai, Ginger Stickney, and Fan-Gang Zeng. Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *The Journal of the Acoustical Society of America*, 123(1):428–440, 2008.

[75] Luxand FaceSDK Documentation. Luxand facesdk 4.0 face detection and recognition library. *Developer s Guide, Copyright*, 2005.

[76] Gail S Donaldson, Heather A Kreft, and Leonid Litvak. Place-pitch discrimination of single-versus dual-electrode stimuli by cochlear implant users a. *The Journal of the Acoustical Society of America*, 118(2):623–626, 2005.

[77] Michael F. Dorman, Julie Liss, Shuai Wang, Visar Berisha, Cimarron Ludwig, and Sarah Cook Natale. Experiments on auditory-visual perception of sentences by users of unilateral, bimodal, and bilateral cochlear implants. *Journal of Speech, Language, and Hearing Research*, 59(6):1505–1519, 2016.

[78] Michael F Dorman and Philipos C Loizou. The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels. *Ear and hearing*, 19(2):162–166, 1998.

[79] Michael F Dorman, Philipos C Loizou, and Dawne Rainey. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *The Journal of the Acoustical Society of America*, 102(4):2403–2411, 1997.

[80] Richard C Dowell, LFA Martin, YC Tong, Graeme M Clark, PM Seligman, and JF Patrick. A 12-consonant confusion study on a multiple-channel cochlear implant patient. *Journal of Speech, Language, and Hearing Research*, 25(4):509–516, 1982.

[81] Richard Drake, A Wayne Vogl, and Adam WM Mitchell. *Gray's anatomy for students*. Elsevier Health Sciences, 2009.

[82] Thomas Eby. Otology, neurotology, and lateral skull base surgery, 2012.

[83] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: An animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graph.*, 35(4):127:1–127:11, July 2016.

[84] Olov Engwall, Olle Bälter, Anne-Marie Öster, and Hedvig Kjellström. Designing the user interface of the computer-based speech training system artur based on early user tests. *Behaviour and Information Technology*, 25(4):353–365, 2006.

[85] Norman P Erber. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech, Language, and Hearing Research*, 12(2):423–425, 1969.

[86] Norman P Erber. *Auditory training.* Alexander Graham Bell Association for the Deaf Washington, DC, 1982.

[87] Tony Ezzat, Gadi Geiger, and Tomaso Poggio. *Trainable videorealistic speech animation*, volume 21. ACM, 2002.

[88] Sascha Fagel and Katja Madany. A 3-d virtual head as a tool for speech therapy for children. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2643–2646, 2008.

[89] Cletus G Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11(4):796–804, 1968.

[90] Michael Fitzpatrick, Jeesun Kim, and Chris Davis. The effect of seeing the interlocutor on auditory and visual speech production in noise. In *Auditory-Visual Speech Processing (AVSP)*, 2011.

[91] Michael Fitzpatrick, Jeesun Kim, and Chris Davis. The intelligibility of lombard speech: communicative setting matters. In *The Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[92] Michael Fitzpatrick, Jeesun Kim, and Chris Davis. The effect of seeing the interlocutor on auditory and visual speech production in noise. *Speech Communication*, 74:37–51, 2015.

[93] James E Flege. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 92:233–277, 1995.

[94] Alexander L Francis and Howard C Nusbaum. Effects of intelligibility on working memory demand for speech perception. *Attention, Perception, and Psychophysics*, 71(6):1360–1374, 2009.

[95] Kai-Ming G Fu, Taylor A Johnston, Ankoor S Shah, Lori Arnold, John Smiley, Troy A Hackett, Preston E Garraghty, and Charles E Schroeder. Auditory cortical neurons respond to somatosensory stimulation. *Journal of Neuroscience*, 23(20):7510–7515, 2003.

[96] Qian-Jie Fu. Angelsim™ cochlear implant hearing loss simulator (version 1.05.05) 2012. http://www.tigerspeech.com/angelsim.

[97] Qian-Jie Fu, John Galvin, Xiaosong Wang, and Geraldine Nogaki. Moderate auditory training can improve speech performance of adult cochlear implant patients. *Acoustics Research Letters Online*, 6(3):106–111, 2005.

[98] Qian-Jie Fu and John J Galvin. Perceptual learning and auditory training in cochlear implant recipients. *Trends in Amplification*, 11(3):193–205, 2007.

[99] Waka Fujisaki, Shinsuke Shimojo, Makio Kashino, and Shin'ya Nishida. Recalibration of audiovisual simultaneity. *Nature neuroscience*, 7(7):773–778, 2004.

[100] Maëva Garnier, Lucie Bailly, Marion Dohen, Pauline Welby, and Hélène Loevenbruck. An acoustic and articulatory study of lombard speech: Global effects on the utterance. In *The Annual International Conference on Spoken Language Processing (INTERSPEECH)*, pages p–2246, 2006.

[101] Maëva Garnier, Lucie Bailly, Marion Dohen, Pauline Welby, and Hélène Lœvenbruck. An acoustic and articulatory study of Lombard speech: Global effects on the utterance. In *The Annual International Conference on Spoken Language Processing (INTERSPEECH)*, 2006.

[102] Maëva Garnier, Nathalie Henrich, and Daniele Dubois. Influence of sound immersion and communicative interaction on the lombard effect. *Journal of Speech, Language, and Hearing Research*, 53(3):588–608, 2010.

[103] Maëva Garnier, Lucie Ménard, and Gabrielle Richard. Effect of being seen on the production of visible speech cues. a pilot study on lombard speech. In *The Annual International Conference on Spoken Language Processing (INTERSPEECH)*, pages 611–614, 2012.

[104] Maëva Garnier, Joe Wolfe, Nathalie Henrich, and John Smith. Interrelationship between vocal effort and vocal tract acoustics: a pilot study. In *The Annual International Conference on Spoken Language Processing (INTERSPEECH)*, pages 2302–2305, 2008.

[105] Christian Gaser and Gottfried Schlaug. Brain structures differ between musicians and non-musicians. *Journal of Neuroscience*, 23(27):9240–9245, 2003.

[106] Isabelle B Gat and Robert W Keith. An effect of linguistic experience: Auditory word discrimination by native and non-native speakers of english. *Audiology*, 17(4):339–345, 1978.

[107] Timothy Q Gentner and Daniel Margoliash. Neuronal populations and single cells representing learned auditory objects. *Nature*, 424(6949):669–674, 2003.

[108] Asif A Ghazanfar and Charles E Schroeder. Is neocortex essentially multisensory? *Trends in cognitive sciences*, 10(6):278–285, 2006.

[109] Theodoros Giannakopoulos. A method for silence removal and segmentation of speech signals, implemented in matlab. *University of Athens, Athens*, 2, 2009.

[110] Theodoros Giannakopoulos and Aggelos Pikrakis. *Introduction to Audio Analysis: A MATLAB Approach*. Academic Press, 2014.

[111] Eleanor Jack Gibson. *Principles of perceptual learning and development.* Appleton-Century-Crofts, 1969.

[112] B. Gick, I. Wilson, and D. Derrick. *Articulatory Phonetics*. Wiley, 2012.

[113] H.J. Giegerich. *English Phonology: An Introduction*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1992.

[114] Charles D Gilbert, Mariano Sigman, and Roy E Crist. The neural basis of perceptual learning. *Neuron*, 31(5):681–697, 2001.

[115] Chris A Glasbey and Kanti V Mardia. A review of image-warping methods. *Journal of applied statistics*, 25(2):155–171, 1998.

[116] Elizabeth Godoy, Maria Koutsogiannaki, and Yannis Stylianou. Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles. *Computer Speech and Language*, 28(2):629–647, 2014.

[117] Robert L Goldstone. Perceptual learning. *Annual review of psychology*, 49(1):585–612, 1998.

[118] Ardeshir Goshtasby. Piecewise linear mapping functions for image registration. *Pattern Recognition*, 19(6):459–466, 1986.

[119] Taro Goto, Marc Escher, Christian Zanardi, and Nadia Magnenat-Thalmann. *MPEG-4 based animation with face feature tracking.* Springer, 1999.

[120] Julie Marie Gottselig, Daniel Brandeis, Gilberte Hofer-Tinguely, Alexander A Borbély, and Peter Achermann. Human central auditory plasticity associated with tone sequence learning. *Learning and Memory*, 11(2):162–171, 2004.

[121] Jeremy Grynpas, Rachel Baker, and Valerie Hazan. *Clear speech strategies and speech perception in adverse listening conditions.* In *International Congress of Phonetic Science (ICPhS)*, 2011.

[122] Abdulrahman Hagr, Soha N Garadat, Sabah M Hassan, Khalid Malki, Yousef Al Ohali, Najwa Al Ghamdi, Abeer Al Nafjan, Ayna Al Masaad, and Sara Al Hamid. The effect of the arabic computer rehabilitation program "rannan" on sound detection and discrimination in children with cochlear implants. *Journal of the American Academy of Audiology*, 27(5):380–387, 2016.

[123] Eric Hamilton. Jpeg file interchange format. *C-Cube Microsystems.*

[124] William J Hardcastle and Nigel Hewlett. *Coarticulation: Theory, data and techniques.* Cambridge University Press, 2006.

[125] Debra M Hardison. Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, 24(4):495–522, 2003.

[126] Michael S Harris, Natalie R Capretta, Shirley C Henning, Laura Feeney, Mark A Pitt, and Aaron C Moberly. Postoperative rehabilitation strategies used by adults with cochlear implants: a pilot study. *Laryngoscope investigative otolaryngology*, 1(3):42–48, 2016.

[127] David JC Hawkey, Sygal Amitay, and David R Moore. Early and rapid perceptual learning. *Nature neuroscience*, 2004.

[128] Valerie Hazan and Rachel Baker. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions a). *The Journal of the Acoustical Society of America*, 130(4):2139–2152, 2011.

[129] Valerie Hazan and Jeesun Kim. Acoustic and visual adaptations in speech produced to counter adverse listening conditions. In *Auditory-Visual Speech Processing (AVSP)*, 2013.

[130] Valerie Hazan, Jeesun Kim, and Yuchun Chen. Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication*, 52(11):996–1009, 2010.

[131] Valerie Hazan, Anke Sennema, Andrew Faulkner, Marta Ortega-Llebaria, Midori Iba, and Hyunsong Chung. The use of visual cues in the perception of non-native consonant contrasts a. *The Journal of the Acoustical Society of America*, 119(3):1740–1751, 2006.

[132] Valerie Hazan, Anke Sennema, Midori Iba, and Andrew Faulkner. Effect of audiovisual perceptual training on the perception and production of consonants by japanese learners of english. *Speech Communication*, 47(3):360–378, 2005.

[133] Valerie Hazan and Andrew Simpson. Cue-enhancement strategies for natural vcv and sentence materials presented in noise. *Speech, Hearing and Language-Work in Progress, Phonetics and Linguistics, University College London*, 9:43–55, 1996.

[134] Karen S Helfer and Richard L Freyman. The role of visual speech cues in reducing energetic and informational masking. *The Journal of the Acoustical Society of America*, 117(2):842–849, 2005.

[135] Alexis Hervais-Adelman, Matthew H Davis, Ingrid S Johnsrude, and Robert P Carlyon. Perceptual learning of noise vocoded words: effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, 34(2):460, 2008.

[136] Sarah Hodkinson. More from music: developing music rehabilitation resources for cochlear implant users. *Counterpoints*, 2014.

[137] Philip Hoole and Marianne Pouplier. Öhman returns: New horizons in the collection and analysis of imaging data in speech production research. *Computer Speech and Language*, 45(Supplement C):253 – 277, 2017.

[138] Jonathan C Horton, Manfred Fahle, Theo Mulder, and Susanne Trauzettel-Klosinski. Adaptation, perceptual learning, and plasticity of brain functions. *Graefe's Archive for Clinical and Experimental Ophthalmology*, pages 1–13, 2017.

[139] Dong-Yan Huang, Susanto Rahardja, and Ee Ping Ong. Lombard effect mimicking. In *Speech Synthesis Workshop (SSW)*, pages 258–263, 2010.

[140] Jessica E Huber and Bharath Chandrasekaran. Effects of increasing sound pressure level on lip and jaw movement parameters and consistency in young adults. *Journal of Speech, Language, and Hearing Research*, 49(6):1368–1379, 2006.

[141] Larry E Humes et al. Understanding the speech-understanding problems of the hearing impaired. *Journal of the American Academy of Audiology*, 2(2):59–69, 1991.

[142] RWG Hunt. Colour science: concepts and methods, quantitative data and formulas. *Journal of Modern Optics*, 15(2):197–197, 1968.

[143] Amir Hussain, Jon Barker, Ricard Marxer, Ahsan Adeel, William Whitmer, Roger Watt, and Peter Drleth. Towards mulit-modal hearing aid design and evaluation in realistic audio-visual setting: Challenges and opportunists. In *1st International Workshop on Challenges in Hearing Assistive Technology (CHAT)*, 2017.

[144] Khaled Huthaily. *Contrastive phonological analysis of Arabic and English*. PhD thesis, The University of Montana, 2003.

[145] Singular Inversions. Facegen modeller (version 3.3)[computer software]. *Toronto, ON: Singular Inversions*, 2008.

[146] Amy Irwin, Michael Pilling, and Sharon M Thomas. An analysis of british regional accent and contextual cue effects on speechreading performance. *Speech Communication*, 53(6):807–817, 2011.

[147] Pamela L Jackson, Allen A Montgomery, and Carl A Binnie. Perceptual dimensions underlying vowel lipreading performance. *Journal of Speech and Hearing Research*, 19(4):796–812, 1976.

[148] Janet Jeffers and Margaret Barley. *Speechreading (lipreading)*. Thomas Springfield, 1971.

[149] Keith Johnson. Speaker normalization in speech perception. *The handbook of speech perception*, pages 363–389, 2005.

[150] Timothy R Jordan, Maxine V Mccotter, and Sharon M Thomas. Visual and audiovisual speech perception with color and gray-scale facial images. *Perception and psychophysics*, 62(7):1394–1404, 2000.

[151] J-C Junqua, Steven Fincke, and Ken Field. The lombard effect: A reflex to better communicate with others in noise. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 2083–2086. IEEE, 1999.

[152] Jean-Claude Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93(1):510–524, 1993.

[153] Adam R Kaiser, Karen Iler Kirk, Lorin Lachs, and David B Pisoni. Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research*, 46(2):390–404, 2003.

[154] H. Kaplan, S.J. Bally, and C. Garretson. *Speechreading: A Way to Improve Understanding*. Gallaudet University Press, 1985.

[155] Uma R Karmarkar and Dean V Buonomano. Temporal specificity of perceptual learning in an auditory discrimination task. *Learning and Memory*, 10(2):141–147, 2003.

[156] Tetsuaki Kawase, Shuichi Sakamoto, Yoko Hori, Atsuko Maki, Yôiti Suzuki, and Toshimitsu Kobayashi. Bimodal audio–visual training enhances auditory adaptation process. *Neuroreport*, 20(14):1231–1234, 2009.

[157] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[158] Jan Kiefer, Steffen Hohl, Ekkehard Stürzebecher, Thomas Pfennigdorff, and Wolfgang Gstöettner. Comparison of speech recognition with different speech coding strategies (speak, cis, and ace) and their relationship to telemetric

measures of compound action potentials in the nucleus ci 24m cochlear implant system. *Audiology*, 40(1):32–42, 2001.

[159] J Kiesslin, MK Pichora-Fuller, S Gatehouse, D Stephens, S Arlinger, TH Chisholm, A Davis, NP Erber, L Hickson, AE Holmes, U Rosenhal, and H von Wedeln. Candidature for and delivery of audiological services: special needs of older people. *Int J Audiol*, 42(2):S92–92S10T, 2003.

[160] Jeehyoung Kim and Wonshik Shin. How to do random allocation (randomization). *Clinics in orthopedic surgery*, 6(1):103–109, 2014.

[161] Jeesun Kim and Chris Davis. Comparing the consistency and distinctiveness of speech produced in quiet and in noise. *Computer Speech and Language*, 28(2):598–606, 2014.

[162] Jeesun Kim and Chris Davis. How visual timing and form information affect speech and non-speech processing. *Brain and language*, 137:86–90, 2014.

[163] Jeesun Kim, Chris Davis, Guillaume Vignali, and Harold Hill. A visual concomitant of the lombard reflex. In *Auditory Visual Speech Processing (AVSP)*, pages 17–22, 2005.

[164] Jeesun Kim, Amanda Sironic, and Chris Davis. Hearing speech in noise: Seeing a loud talker is better. *Perception*, 40(7):853–862, 2011.

[165] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[166] Yoko Kitano, Bruce M Siegenthaler, and Richard G Stoker. Facial hair as a factor in speechreading performance. *Journal of communication disorders*, 18(5):373–381, 1985.

[167] V. Kitanovski and E. Izquierdo. Augmented reality mirror for virtual facial alterations. In *18th IEEE International Conference on Image Processing*, pages 1093–1096, Sept 2011.

[168] Dennis H Klatt. Review of the arpa speech understanding project. *The Journal of the Acoustical Society of America*, 62(6):1345–1366, 1977.

[169] Dawn Burton Koch, Mary Joe Osberger, Phil Segel, and Dorcas Kessler. Hiresolutiontm and conventional sound processing in the hiresolutiontm bionic ear: using appropriate outcome measures to assess speech recognition ability. *Audiology and Neurotology*, 9(4):214–223, 2004.

[170] Nina Kraus, Therese McGee, Thomas D Carrell, Cynthia King, Kelly Tremblay, and Trent Nicol. Central auditory system plasticity associated with speech discrimination training. *Journal of cognitive neuroscience*, 7(1):25–32, 1995.

[171] NCSU Phonology Lab. Penn phonetics lab forced aligner (p2fa). http://phon.chass.ncsu.edu/cgi-bin/step7.cgi.

[172] Lorin Lachs, David B Pisoni, and Karen Iler Kirk. Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: A first report. *Ear and hearing*, 22(3):236, 2001.

[173] Karen Lander and Cheryl Capek. Investigating the impact of lip visibility and talking style on speechreading performance. *Speech Communication*, 55(5):600–605, 2013.

[174] Simon Landry, Justine Lévesque, and François Champoux. Breaking news: Brain plasticity an obstacle for cochlear implant rehabilitation. *The Hearing Journal*, 65(8):26–28, 2012.

[175] Harlan Lane. Foreign accent and speech distortion. *The Journal of the Acoustical Society of America*, 35(4):451–453, 1963.

[176] Harlan Lane and Bernard Tranel. The lombard sign and the role of hearing in speech. *Journal of Speech, Language, and Hearing Research*, 14(4):677–709, 1971.

[177] Maria Luisa Garcia Lecumberri, Martin Cooke, and Anne Cutler. Non-native speech perception in adverse conditions: A review. *Speech Communication*, 52(11):864–886, 2010.

[178] ML Garcia Lecumberri and Martin Cooke. Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America*, 119(4):2445–2454, 2006.

[179] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas S Huang. Avicar: audio-visual speech corpus in a car environment. In *The Annual International Conference on Spoken Language Processing (INTERSPEECH)*, 2004.

[180] Yuencheng Lee, Demetri Terzopoulos, and Keith Waters. Realistic modeling for facial animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62. ACM, 1995.

[181] R. Leonard. A database for speaker-independent digit recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 9, pages 328–331, Mar 1984.

[182] Howard Levene. Robust tests for equality of variances. *Contributions to probability and statistics*, 1:278–292, 1960.

[183] Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. Data-driven enhancement of facial attractiveness. *ACM Transactions on Graphics (TOG)*, 27(3):38, 2008.

[184] Wu Li, Valentin Piëch, and Charles D Gilbert. Perceptual learning and top-down influences in primary visual cortex. *Nature neuroscience*, 7(6):651, 2004.

[185] Xing Li, Kaibao Nie, Nikita S Imennov, Jay T Rubinstein, and Les E Atlas. Improved perception of music with a harmonic based algorithm for cochlear implants. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(4):684–694, 2013.

[186] Alvin M Liberman and Ignatius G Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985.

[187] B Lindblom. Explaining phonetic variation: A sketch of the handh theory. *Speech Production and Speech Modelling*, 55:403, 2012.

[188] Ce Liu, Antonio Torralba, William T Freeman, Frédo Durand, and Edward H Adelson. Motion magnification. *ACM transactions on graphics (TOG)*, 24(3):519–526, 2005.

[189] Scott E Lively, John S Logan, and David B Pisoni. Training japanese listeners to identify english/r/and/l/: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, 94(3):1242–1255, 1993.

[190] Anders Löfqvist, Birgitta Sahlén, and Tina Ibertsson. Vowel spaces in swedish adolescents with cochlear implants. *The Journal of the Acoustical Society of America*, 128(5):3064–3069, 2010.

[191] John S Logan, Scott E Lively, and David B Pisoni. Training japanese listeners to identify english/r/and/l: A first report. *The Journal of the Acoustical Society of America*, 89(2):874–886, 1991.

[192] Etienne Lombard. Le signe de l'elevation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37(101-119):25, 1911.

[193] Youyi Lu. *Production and perceptual analysis of speech produced in noise*. PhD thesis, University of Sheffield, 2010.

[194] Youyi Lu and Martin Cooke. Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America*, 124(5):3261–3275, 2008.

[195] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *The International Joint Conference in Artificial Intelligence (IJCAI)*. Vancouver, BC, Canada, 1981.

[196] Wei Ji Ma, Xiang Zhou, Lars A Ross, John J Foxe, and Lucas C Parra. Lip-reading aids word recognition most in moderate noise: a bayesian explanation using high-dimensional feature space. *PLoS One*, 4(3):e4638, 2009.

[197] David JC MacKay. Innovation and intellectual property rights. In *Information theory, inference and learning algorithms*, chapter 20. Cambridge university press, 2003.

[198] Alison MacLeod and Quentin Summerfield. Quantifying the contribution of vision to speech perception in noise. *British journal of audiology*, 21(2):131–141, 1987.

[199] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[200] Eleanor A Maguire, Katherine Woollett, and Hugo J Spiers. London taxi drivers and bus drivers: a structural mri and neuropsychological analysis. *Hippocampus*, 16(12):1091–1101, 2006.

[201] MJ Manrique, JM Espinosa, A Huarte, M Molina, R Garcia-Tapia, and J Artieda. Cochlear implants in post-lingual persons: results during the first five years of the clinical course. *Acta otorrinolaringologica espanola*, 49(1):19–24, 1998.

[202] Dominic Massaro. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Lawrence Erlbaum Associates, Inc, 1987.

[203] Dominic W Massaro. A computer-animated tutor for spoken and written language learning. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 172–175. ACM, 2003.

[204] Dominic W Massaro and Michael M Cohen. Phonological context in speech perception. *Perception and psychophysics*, 34(4):338–348, 1983.

[205] Dominic W Massaro and Michael M Cohen. Perception of synthesized audible and visible speech. *Psychological Science*, 1(1):55–63, 1990.

[206] Dominic W Massaro and David G Stork. Speech recognition and sensory integration: a 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist*, pages 236–244, 1998.

[207] Iain Matthews, Timothy F Cootes, J Andrew Bangham, Stephen Cox, and Richard Harvey. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213, 2002.

[208] Sven L Mattys, Matthew H Davis, Ann R Bradlow, and Sophie K Scott. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7-8):953–978, 2012.

[209] Lynn Hansberry Mayo, Mary Florentine, and Søren Buus. Age of second-language acquisition and perception of speech in noise. *Journal of speech, language, and hearing research*, 40(3):686–693, 1997.

[210] Matthew McGrath. *An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces.* PhD thesis, University of Nottingham, 1985.

[211] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[212] Diane Meador, James E Flege, and Ian RA Mackay. Factors affecting the recognition of words in a second language. *Bilingualism: Language and Cognition*, 3(1):55–67, 2000.

[213] Andrea Mechelli, Jenny T Crinion, Uta Noppeney, John O'doherty, John Ashburner, Richard S Frackowiak, and Cathy J Price. Neurolinguistics: structural plasticity in the bilingual brain. *Nature*, 431(7010):757–757, 2004.

[214] Stefano Melacci, Lorenzo Sarti, Marco Maggini, and Marco Gori. A template-based approach to automatic face enhancement. *Pattern Analysis and Applications*, 13(3):289–300, 2010.

[215] Hans Menning, Larry E Roberts, and Christo Pantev. Plastic changes in the auditory cortex induced by intensive frequency discrimination training. *Neuroreport*, 11(4):817–822, 2000.

[216] M Alex Meredith and Barry E Stein. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J Neurophysiol*, 56(3):640–662, 1986.

[217] Michael M Merzenich. Temporal processing deficits of language-learning. *Proc. Natl. Acad. Sci. USA*, 90:9135, 1993.

[218] Michael M Merzenich, JH Kaas, J Wall, RJ Nelson, M Sur, and D Felleman. Topographic reorganization of somatosensory cortical areas 3b and 1 in adult monkeys following restricted deafferentation. *Neuroscience*, 8(1):33–55, 1983.

[219] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference*

on audio and video-based biometric person authentication, volume 964, pages 965–966, 1999.

[220] Pascale Michelon. Brain plasticity: How learning changes your brain. *Sharpbrains*, 2008.

[221] MJ Middelweerd and R Plomp. The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America*, 82(6):2145–2147, 1987.

[222] George A Miller and Patricia E Nicely. An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352, 1955.

[223] Hansjörg Mixdorff, Ulrich Pech, Chris Davis, and Jeesun Kim. Map task dialogs in noise–a paradigm for examining lombard speech. In *International Congress of Phonetic Science (ISPhS)*, pages 1329–1332, 2007.

[224] Allen A Montgomery and Pamela L Jackson. Physical characteristics of the lips underlying vowel lipreading performance. *The Journal of the Acoustical Society of America*, 73(6):2134–2144, 1983.

[225] Brian Moore, Lorraine Tyler, and William Marslen-Wilson. *The perception of speech: from sound to meaning*. Oxford University Press, 2009.

[226] Brian CJ Moore. *Cochlear hearing loss: physiological, psychological and technical issues*. John Wiley and Sons, 2007.

[227] Roger K Moore and Mauro Nicolao. Reactive speech synthesis: Actively managing phonetic contrast along an handh continuum. In *17th international congress of phonetics sciences (ICPhS)*, 2011.

[228] Thomas J Moore. Voice communications jamming research. In *Advisory Group for Aerospace Research and Development Conference*, 1981.

[229] Masahiro Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.

[230] Julia A Mossbridge, Matthew B Fitzgerald, Erin S O'Connor, and Beverly A Wright. Perceptual-learning evidence for separate processing of asynchrony and order tasks. *Journal of Neuroscience*, 26(49):12708–12716, 2006.

[231] MPEG-4. Mpeg-4 international standard, 1997. http://mpeg.chiariglione.org/standards/mpeg-4/mpeg-4.html.

[232] Alexandre Nentchev. *Numerical analysis and simulation in microelectronics by vector finite elements*. PhD thesis, Institute for Microelectronics at the TU Wien, 2008.

[233] Arlene C Neuman. Central auditory system plasticity and aural rehabilitation of adults. *Journal of rehabilitation research and development*, 42(4):169, 2005.

[234] Anderson Jonas das Neves, Ana Claudia Moreira Almeida Verdu, Adriane de Lima MortariMoret, and Leandra Tabanez do Nascimento Silva. The implications of the cochlear implant for development of language skills: a literature review. *Speech, Language, Hearing Sciences and Education Journal (Revista CEFAC)*, 17(5):1643–1656, 2015.

[235] Kevin Nguyen and Shweta Panditrao. *Mobile Audiometry Application*. PhD thesis, Santa Clara: Santa Clara University, 2014., 2014.

[236] Mauro Nicolao, Javier Latorre, and Roger K Moore. C2h: A computational model of handh-based phonetic contrast in synthetic speech. In *The Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[237] Kaibao Nie, Amy Barco, and Fan-Gang Zeng. Spectral and temporal cues in cochlear implant speech perception. *Ear and hearing*, 27(2):208–217, 2006.

[238] Non-native. Oxford dictionaries. Oxford University Press, 2017.

[239] HC Nusbaum and JS Magnuson. Talker normalization: Phonetic constancy as a cognitive process. *Talker variability in speech processing*, pages 109–132, 1997.

[240] Marta Ortega-Llebaria, Andrew Faulkner, and Valerie Hazan. Auditory-visual l2 speech perception: Effects of visual cues and acoustic-phonetic context for spanish learners of english. In *Auditory-Visual Speech Association (AVSP)*, 2001.

[241] D. O'Shaughnessy. Linear predictive coding. *IEEE Potentials*, 7(1):29–32, Feb 1988.

[242] Christo Pantev, Bernhard Ross, Takako Fujioka, Laurel J Trainor, Michael Schulte, and Matthias Schulz. Music and learning-induced cortical plasticity. *Annals of the New York Academy of Sciences*, 999(1):438–450, 2003.

[243] Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci, and J Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages II–2017. IEEE, 2002.

[244] Jonathan E Peelle and Mitchell S Sommers. Prediction and constraint in audiovisual speech perception. *Cortex*, 68:169–181, 2015.

[245] Catherine Pelachaud, Norman I Badler, and Mark Steedman. Linguistic issues in facial animation. In *Computer animation'91*, pages 15–30. Springer, 1991.

[246] Nathaniel R Peterson, David B Pisoni, and Richard T Miyamoto. Cochlear implants and spoken language processing abilities: Review and assessment of the literature. *Restorative neurology and neuroscience*, 28(2):237–250, 2010.

[247] Benjamin Picart, Thomas Drugman, and Thierry Dutoit. Analysis and synthesis of hypo-and hyperarticulated speech. In *Speech Synthesis Workshop (SSW)*, pages 270–275, 2010.

[248] James O Pickles. *An introduction to the physiology of hearing*, volume 2. Academic press London, 1988.

[249] Stéphane Pigeon. Online hearing test and audiogram printout. https://hearingtest.online/.

[250] Michael Pilling and Sharon Thomas. Audiovisual cues and perceptual learning of spectrally distorted speech. *Language and speech*, 54(4):487–497, 2011.

[251] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[252] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004.

[253] Gilles Pourtois, Karsten S Rauss, Patrik Vuilleumier, and Sophie Schwartz. Effects of perceptual learning on primary visual cortex activity in humans. *Vision research*, 48(1):55–62, 2008.

[254] Lawrence R Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. PTR Prentice Hall, 1993.

[255] Charles M Rader. *Digital Processing of Speech Signals. Signal Processing Series*. JSTOR, 1980.

[256] Vilayanur S Ramachandran. *Perception of shape from shading*. Nature Publishing Group, 1988.

[257] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[258] Karen S Reinke, Yu He, Chenghua Wang, and Claude Alain. Perceptual learning modulates sensory evoked response during vowel segregation. *Cognitive Brain Research*, 17(3):781–791, 2003.

[259] Anne-Raphaëlle Richoz, Rachael E Jack, Oliver GB Garrod, Philippe G Schyns, and Roberto Caldara. Reconstructing dynamic mental models of facial expressions in prosopagnosia reveals distinct representations for identity and expression. *Cortex*, 65:50–64, 2015.

[260] Jordi Robert-Ribes, Jean-Luc Schwartz, Tahar Lallouache, and Pierre Escudier. Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of french oral vowels in noise. *The Journal of the Acoustical Society of America*, 103(6):3677–3689, 1998.

[261] Catherine L Rogers, Jennifer J Lister, Dashielle M Febo, Joan M Besing, and Harvey B Abrams. Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, 27(03):465–485, 2006.

[262] M. Rogers, K. Walker, T.G. Williams, M.D.L. Gorce, and M. Tosas. Building systems for tracking facial features across individuals and groups, August 18 2015. US Patent 9,111,134.

[263] William Rogers. The History of English Phonemes. Furman University, http://facweb.furman.edu/ wrogers/phonemes, 2000.

[264] Stuart Rosen, Andrew Faulkner, and Lucy Wilkinson. Perceptual adaptation by normal listeners to upward shifts of spectral information in speech and its relevance for users of cochlear implants. *Journal of the Acoustical Society of America*, 106:3629–3636, 1999.

[265] Lawrence D Rosenblum. Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, 17(6):405–409, 2008.

[266] Lawrence D Rosenblum, Jennifer A Johnson, and Helena M Saldana. Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research*, 39(6):1159–1170, 1996.

[267] Lawrence D Rosenblum and Helena M Saldaña. An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 22(2):318, 1996.

[268] J Amiel Rosenkranz, Holly Moore, and Anthony A Grace. The prefrontal cortex regulates lateral amygdala neuronal plasticity and responses to previously conditioned stimuli. *The Journal of Neuroscience*, 23(35):11054–11064, 2003.

[269] Lars A Ross, Dave Saint-Amour, Victoria M Leavitt, Daniel C Javitt, and John J Foxe. Do you see what i am saying? exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, 17(5):1147–1153, 2006.

[270] M Ross. Is auditory training effective in improving listening skills. *Hear Loss Magazine*, pages 25–27, 2011.

[271] Julien Rouger, Bernard Fraysse, Olivier Deguine, and Pascal Barone. Mcgurk effects in cochlear-implanted deaf subjects. *Brain research*, 1188:87–99, 2008.

[272] DG Russell. *Spatial location cues and movement production.* Academic Press New York, 1976.

[273] Helena M Saldaña and Lawrence D Rosenblum. Visual influences on auditory pluck and bow judgments. *Perception and psychophysics*, 54(3):406–416, 1993.

[274] Mikko Sams, Reijo Aulanko, Matti Hämäläinen, Riitta Hari, Olli V Lounasmaa, Sing-Teh Lu, and Juha Simola. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience letters*, 127(1):141–145, 1991.

[275] Conrad Sanderson. The vidtimit database. Technical report, IDIAP. Available: https://infoscience.epfl.ch/record/82748/files/com02-06.pdf, 2002.

[276] Jared Sanson and Richard Green. Face replacement demo using the kinect depth sensor. Technical report, Tech. Rep.[Online]. Available: http://jared. geek. nz/face-replace/files/COSC42820ComputerVision-FaceReplace. pdf.

[277] Inga M Schepers, Daniel Yoshor, and Michael S Beauchamp. Electrocorticography reveals enhanced visual cortex responses to visual speech. *Cerebral Cortex*, 25(11):4103–4110, 2014.

[278] Efrat A Schorr, Nathan A Fox, Virginie van Wassenhove, and Eric I Knudsen. Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18748–18750, 2005.

[279] V.M. Scott. *Belonging*. Flying Fingers Club Series. Kendall Green Publications, Gallaudet University Press, 1987.

[280] Kaoru Sekiyama. Differences in auditory-visual speech perception between japanese and americans: Mcgurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E)*, 15(3):143–158, 1994.

[281] Robert V Shannon, Qian-Jie Fu, John Galvin, and Lendra Friesen. Speech perception with cochlear implants. In *Cochlear implants: auditory prostheses and electric hearing*, pages 334–376. Springer, 2004.

[282] Robert V Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303, 1995.

[283] Juraj Simko, Stefan Benu, and Martti Vainio. Hyperarticulation in lombard speech: A preliminary study. In *Proceedings of the 7th international conference on Speech Prosody*, 2014.

[284] Juraj Šimko, Štefan Beňuš, and Martti Vainio. Hyperarticulation in lombard speech: Global coordination of the jaw, lips and the tongue. *The Journal of the Acoustical Society of America*, 139(1):151–162, 2016.

[285] Adrian P Simpson. Phonetic differences between male and female speech. *Language and Linguistics Compass*, 3(2):621–640, 2009.

[286] Mark D Skowronski and John G Harris. Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication*, 48(5):549–558, 2006.

[287] Irwin Sobel and Gary Feldman. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, pages 271–272, 1968.

[288] Mitchell S Sommers, Nancy Tye-Murray, and Brent Spehar. Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and hearing*, 26(3):263–275, 2005.

[289] Pamela Souza and Stuart Rosen. Effects of envelope bandwidth on the intelligibility of sine-and noise-vocoded speech. *The Journal of the Acoustical Society of America*, 126(2):792–805, 2009.

[290] Ann Spriet, Lieselot Van Deun, Kyriaky Eftaxiadis, Johan Laneau, Marc Moonen, Bas Van Dijk, Astrid Van Wieringen, and Jan Wouters. Speech understanding in background noise with the two-microphone adaptive beamformer beam™ in the nucleus freedom™ cochlear implant system. *Ear and hearing*, 28(1):62–72, 2007.

[291] Paula C Stacey and A Quentin Summerfield. Effectiveness of computer-based auditory training in improving the perception of noise-vocoded speech. *The Journal of the Acoustical Society of America*, 121(5):2923–2935, 2007.

[292] Herman JM Steeneken. The measurement of speech intelligibility. *Proceedings-Institute of Acoustics*, 23(8):69–76, 2001.

[293] Ryan A Stevenson, Sterling W Sheffield, Iliza M Butera, René H Gifford, and Mark T Wallace. Multisensory integration in cochlear implant recipients. *Ear and Hearing*, 2017.

[294] Winifred Strange. Cross-language studies of speech perception: A historical review. *Speech perception and linguistic experience: Issues in cross-language research*, pages 3–45, 1995.

[295] Maren Stropahl and Stefan Debener. Auditory cross-modal reorganization in cochlear implant users indicates audio-visual integration. *NeuroImage: Clinical*, 2017.

[296] William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

[297] Q Summerfield, A MacLeod, M McGrath, and M Brooke. Lips, teeth, and the benefits of lipreading. *Handbook of research on face processing*, pages 223–233, 1989.

[298] Quentin Summerfield. Use of visual information for phonetic perception. *Phonetica*, 36(4-5):314–331, 1979.

[299] Quentin Summerfield. Some preliminaries to a compre hensive account of audio-visual speech perception. *Hearing by eye: The psychology of lip-reading*, pages 3–226, 1987.

[300] Quentin Summerfield and Matthew McGrath. Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology*, 36(1):51–74, 1984.

[301] W. Van Summers, David B. Pisoni, Robert H. Bernacki, Robert I. Pedlow, and Michael A. Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3):917–928, 1988.

[302] Robert Sweetow and Catherine V Palmer. Efficacy of individual auditory training in adults: A systematic review of the evidence. *Journal of the American Academy of Audiology*, 16(7):494–504, 2005.

[303] Robert W Sweetow and Jennifer Henderson Sabes. The need for and development of an adaptive listening and communication enhancement (lace™) program. *Journal of the American Academy of Audiology*, 17(8):538–558, 2006.

[304] Robert W Sweetow and Jennifer Henderson Sabes. Listening and communication enhancement (lace). In *Seminars in Hearing*, volume 28, pages 133–141, 2007.

[305] Robert W Sweetow and Jennifer Henderson Sabes. Technologic advances in aural rehabilitation: Applications and innovative methods of service delivery. *Trends in Amplification*, 11(2):101–111, 2007.

[306] Lisa YW Tang, Beverly Hannah, Allard Jongman, Joan Sereno, Yue Wang, and Ghassan Hamarneh. Examining visible articulatory features in clear and plain speech. *Speech Communication*, 75:1–13, 2015.

[307] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284. Eurographics Association, 2012.

[308] Audacity Team. Audacity (version 2.0. 2). *Audio editor and recorder*, 2012.

[309] Barry Theobald, Richard Harvey, Stephen Cox, G Owen, and C Lewis. Lip-reading enhancement for law enforcement. In *SPIE conference on Optics and Photonics for Counterterrorism and Crime Fighting*, pages 640205–1, 2006.

[310] Marko Tkalcic and Jurij F Tasic. *Colour spaces: perceptual, historical and applicational background*, volume 1. IEEE, 2003.

[311] Miguel Torres and Fernando Giráldez. The development of the vertebrate inner ear. *Mechanisms of development*, 71(1):5–21, 1998.

[312] Kelly Tremblay, Nina Kraus, and Thomas McGee. The time course of auditory perceptual learning: neurophysiological changes during speech-sound training. *Neuroreport*, 9(16):3557–3560, 1998.

[313] KL Tremblay and N Kraus. Beyond the ear: central auditory plasticity. *Otorinolaringol*, 52(3):93–100, 2002.

[314] Fernando Trujillo. Speech production process. University Lecture, University of Granada, 2001.

[315] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.

[316] Nancy Tye-Murray. *Foundations of aural rehabilitation: Children, adults, and their family members*. Cengage learning, 2014.

[317] Richard S Tyler, Holly Fryauf-Bertschy, Danielle MR Kelsay, Bruce J Gantz, George P Woodworth, Aaron Parkinson, et al. Speech perception by prelingually deaf children using cochlear implants. *Otolaryngology-Head and Neck Surgery*, 117(3):180–187, 1997.

[318] Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. Fourier principles for emotion-based human figure animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 91–96. ACM, 1995.

[319] Sander J Van Wijngaarden, Herman JM Steeneken, and Tammo Houtgast. Quantifying the intelligibility of speech in noise for non-native listeners. *The Journal of the Acoustical Society of America*, 111(4):1906–1916, 2002.

[320] Eric Vatikiotis-Bateson, Adriano Vilela Barbosa, Cheuk Yi Chow, Martin Oberg, Johanna Tan, and Hani C. Audiovisual Lombard speech: Yehia. *reconciling production and perception*. Auditory-Visual Speech Processing (AVSP), 2007.

[321] Eric Vatikiotis-Bateson, Victor Chung, Kevin Lutz, Nicole Mirante, Jolien Otten, and Johanna Tan. Auditory, but perhaps not visual, processing of lombard speech. *The Journal of the Acoustical Society of America*, 119(5):3444–3444, 2006.

[322] Antoine Vigneron. Computing a delaunay triangulation. *National of University of Singapore*, 2004.

[323] Jean Vroomen, Mirjam Keetels, Béatrice De Gelder, and Paul Bertelson. Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive brain research*, 22(1):32–35, 2004.

[324] Brian E Walden, Robert A Prosek, Allen A Montgomery, Charlene K Scherr, and Carla J Jones. Effects of training on the visual recognition of consonants. *Journal of Speech, Language, and Hearing Research*, 20(1):130–145, 1977.

[325] Jue Wang, Steven M Drucker, Maneesh Agrawala, and Michael F Cohen. The cartoon animation filter. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 1169–1173. ACM, 2006.

[326] Marilyn D Wang and Robert C Bilger. Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, 54(5):1248–1266, 1973.

[327] Yue Wang, Dawn M Behne, and Haisheng Jiang. Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, 124(3):1716–1726, 2008.

[328] Rachel V Wayne and Ingrid S Johnsrude. The role of visual speech information in supporting perceptual learning of degraded speech. *Journal of Experimental Psychology: Applied*, 18(4):419, 2012.

[329] Eric W. Weisstein. Bézier curve. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/BezierCurve.html, 2012.

[330] Lynn Williams. The production of speech sounds. University Lecture, East Tennessee State University, 2006.

[331] Blake S Wilson, Charles C Finley, Dewey T Lawson, Robert D Wolford, and Mariangeli Zerbi. Design and evaluation of a continuous interleaved sampling (cis) processing strategy for multichannel cochlear implants. *Journal of rehabilitation research and development*, 30:110–110, 1993.

[332] Andrew Witkin and Zoran Popovic. Motion warping. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 105–108. ACM, 1995.

[333] David L Woods and E William Yund. Perceptual training of phoneme identification for hearing loss. In *Seminars in Hearing*, volume 28, pages 110–119, 2007.

[334] Beverly A Wright and Yuxuan Zhang. A review of the generalization of auditory learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1515):301–311, 2009.

[335] Zilong Xie, Han-Gyol Yi, and Bharath Chandrasekaran. Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener. *PloS one*, 9(12):e114439, 2014.

[336] Han-Gyol Yi, Jasmine EB Phelps, Rajka Smiljanic, and Bharath Chandrasekaran. Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America*, 134(5):EL387–EL393, 2013.

[337] S-C Yin, Richard Rose, Oscar Saz, and Eduardo Lleida. A study of pronunciation verification in a speech therapy application. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4609–4612. IEEE, 2009.

[338] Steve J Young and Sj Young. *The HTK hidden Markov model toolkit: Design and philosophy.* University of Cambridge, Department of Engineering, 1993.

[339] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1944–1951. IEEE, 2013.

[340] Yinsheng Zhou, Khe Chai Sim, Patsy Tan, and Ye Wang. Mogat: mobile games with auditory training for children with cochlear implants. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 429–438. ACM, 2012.

[341] Ziheng Zhou, Guoying Zhao, Xiaopeng Hong, and Matti Pietikäinen. A review of recent advances in visual speech decoding. *Image and vision computing*, 32(9):590–605, 2014.

[342] Victor Zue, Stephanie Seneff, and James Glass. Speech database development at mit: Timit and beyond. In Speech Communication,, pages 351–356. 1990.