A forensic phonetic study of the vocal responses of individuals in distress

Lisa S. Roberts

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Language and Linguistic Science

September 2012

Abstract

The production and perception of emotional speech is of growing importance to forensic speech scientists. They are often asked by instructing parties to provide an opinion as to whether recordings representing a violent attack are genuine, and whether speech material reflects real distress. However, they are prohibited from making statements regarding the psychological states of speakers by the International Association of Forensic Phonetics and Acoustics Code of Practice (IAFPA 2004).

This study investigates two principal questions. First, it investigates how distress speech can be manifested acoustically. In so doing it proposes a taxonomy for comparing distress speech across speakers, assists in delimiting the boundaries of the vocal repertoire, and considers the extent to which acoustic measures of distress speech can distinguish between the vocalisations of real victims and actors. Second, it investigates whether listeners can discriminate between genuine and acted distress portrayals, and to what extent familiarity with forensic material increases listeners' ability.

Recordings from authentic criminal cases involving violent attack are compared with re-enactments by trained actors. Acoustic analyses examine F0, intensity, vowel formant frequencies and articulation rate. The recordings are also used as stimuli in a perceptual listening test, comparing the performance of lay listeners, police call takers and forensic practitioners.

The findings lend support to the view that assessments of distress should be exercised with extreme caution. On the one hand, acoustic parameters can distinguish between non-distress and distress conditions, but cannot discriminate between acted and authentic distress, and so IAFPA's refrain from such an assessment is justified. On the other, listeners who are familiar with authentic distress data, such as police call takers and forensic practitioners, are better able to differentiate between acted and authentic distress than lay listeners. Thus, if an assessment were to be made, the forensic practitioners may be the best group to do so.

Table of Contents

List of tables

List of figures

Acknowledgements

As I progressed further through the PhD programme, this was the section I'd been most looking forward to writing. Research of any kind is challenging, frustrating, quite often soul-destroying (or should that be strengthening?) and yet it's also extremely rewarding. I doubt I would have made it through if it hadn't been for the good many people who have been there to support me academically, financially and emotionally over these past few years. However, now that the end is in sight, I find it hard to know where to start thanking them all. I am indebted to many people whose contributions have made this research project possible.

I wish to first thank Dr. Sneddon, whose encouragement during my early academic forays at St. Andrews led me to consider post-graduate study in linguistics and phonetics. Our chats at Macgregor's over assam tea and a slice of cake gave me the confidence to apply for the forensic speech science programme at York.

Studying at York turned out to be a life-changing event. Thanks to Paul Foulkes, Peter French, and Dom Watt, I enjoyed the MSc more than I ever could have imagined. They fostered my interest in forensic speech science, especially speech in distress. It's through their enthusiasm and support that I decided to take the next step of enrolling as a PhD student. Their comments, advice and encouragement have been invaluable throughout the evolution of this thesis. I am so very proud to be able to contribute to this field, if only in a small way.

Thanks must also go to Carmen Llamas. Although not directly involved with my thesis, she's always been on hand for advice and a beer. She trusted me to collect speech samples from across the North East for her own work, and has made me question my own speech patterns on countless times. I'm looking forward to the continuation of the Tees, Wear And Tyne research.

The staff at JPFA have been an absolute pleasure to work with (and on some occasions, for). They've been encouraging, helpful, motivating, and they know how to truly appreciate tea. The many months of analysing (and re-analysing) my data wouldn't have been half as enjoyable if it wasn't for the special JPFA environment.

The York Forensic research group have been great at providing their ears and their thoughts at various stages of my PhD, and even acted as experiment subjects. Thank you for your time and attentive reflections as my project developed.

I consider myself lucky to be part of a department where staff members, even those not directly involved in my thesis topic, have been available for guidance and support. Bill Haddican, Sam Hellmuth and Tamar Keren-Portnoy have provided me with comments and advice on a variety of topics that have cropped up during my thesis. They have also been instrumental in developing my teaching career.

I owe many thanks to the Economic and Social Research Council for funding assistance, and to Paul for helping me get through the application!

Morwenna Rowe, Dan Barnard, and all the actors who generously gave their time and energy to make the drama workshop, and thus my data collection, a success, receive my sincerest thanks. Lynn Stevens and Linda Turvey also proved to be invaluable when it came to recruiting participants for my perceptual experiment, and I am grateful to all those who willingly gave their time to take part and further my research.

My time at York wouldn't have been the same if it weren't for the other postgrads (and RAs) who have been able to share the trials, tribulations, and joys of being a York postgrad. Special thanks to Colleen, Christin, and Erica, my PhD partners in crime, and Jen, my inspiring post-doc friend, for keeping me sane during this time. Your friendship and camaraderie are always appreciated. I hope I am able to offer the same in return.

Jen Nycz and Stuart Brown receive additional thank yous not only for starting many a thought provoking conversation (in linguistics and often beyond), but also for their supreme Catan playing skills. I'd sell you all my wheat if you wouldn't have to go off and live in Brazil/Bath/the US all the time.

My non-linguistic friends have always been close by for emotional support (often in the form of providing much needed tea and cake). There are too many of you list, so I'll thank you in person instead, perhaps over afternoon tea ☺. I am grateful to call you all friends.

My family have been overwhelmingly supportive throughout all my endeavours over the years. They have given me the courage to embrace life and all it throws you, (usually) without regret. My grandparents and godparents, and, above all, my parents, have always been there for me. Thank you.

And finally to Eytan. I don't think you'll ever know how much your love and support means to me. We hadn't yet met when I first started the PhD, but I'm so glad we're together as I finish it. Here's to many more adventures together.

Author's declaration

This is to certify that this thesis comprises original work and that all contributions from external sources have been explicitly stated and referenced appropriately.

I also declare that aspects of the research have been previously presented at conferences and in proceedings papers. These publications are listed as follows:

Publications:

Roberts, L. (2011). Acoustic effects of authentic and acted distress on fundametal frequency and vowel quality. *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong*, 1694-1697.

Roberts, L. (2010). Real and acted responses of distress: an auditory and acoustic analysis of extreme stress and emotion. In Antonis Botinis (Ed.) Exling 2010: *Proceedings of the ISCA Tutorial and Research Workshop on Experimental Linguistics*. Athens, Greece: ISCA and the University of Athens, 149-152.

Roberts, L. & French, J.P. (2010). How context affects perception: judging distress and linguistic content in forensic audio recordings. In Antonis Botinis (Ed.) Exling 2010: *Proceedings of the ISCA Tutorial and Research Workshop on Experimental Linguistics*. Athens, Greece: ISCA and the University of Athens, 153-156.

Presentations:

Roberts, L. (2012). The influence of experience on perceptions of distress. *The 2012 International Association for Forensic Phonetics and Acoustics Annual Conference (IAFPA 2012)*, Santander, Spain.

Roberts, L. (2012). Acoustic cues of distress in actors and victims. *The 2012 Colloquium of the British Association of Academic Phoneticians (BAAP 2012)*, Leeds, UK.

Roberts, L. (2011). Acoustic effects of authentic and acted distress on fundamental frequency and vowel quality. *The 17th International Congress of Phonetic Sciences (ICPHS 2011)*, Hong Kong, China.

Roberts, L. (2011). Acoustic characteristics of distress speech in real vicitms and trained actors. *The 2011 International Association for Forensic Phonetics and Acoustics Annual Conference (IAFPA 2011)*, Vienna, Austria.

Roberts, L. (2010). Real and acted responses of distress: an auditory and acoustic analysis of extreme stress and emotion. *The ISCA Tutorial and Research Workshop on Experimental Linguistic (Exling 2010),* Athens, Greece.

Roberts, L. & French, J.P. (2010). How context affects perception: judging distress and linguistic content in forensic audio recordings. *The ISCA Tutorial and Research Workshop on Experimental Linguistic (Exling 2010),* Athens, Greece.

Roberts, L. (2010). Authentic and Acted responses to physical and emotion distress . *Sâo Paulo School of Advanced Studies in Speech Dynamics Dinâmica Fônica Workshop (Dinafon)*, Sâo Paulo, Brazil.

Roberts, L., French, J. P., & Harrison, P. (2010).The Influence of Context on Listeners' Perceptions of Distress and Linguistic Content in Forensic Audio Recordings. *The 2010 Colloquium of the British Association of Academic Phoneticians (BAAP 2010),* London, UK.

Roberts, L. (2010). Phonetic Cues of Psychologically Stressed Individuals in Forensic Contexts. *The 2010 Colloquium of the British Association of Academic Phoneticians (BAAP 2010),* London, UK.

Roberts, L. (2009). Phonetic Characteristics of Real and Simulated Responses to Violent Attacks. *The 2009 International Association for Forensic Phonetics and Acoustics Annual Conference (IAFPA 2009)*, Cambridge, UK.

Signed:……  …… (candidate)

Date: ……02 December 2014……

1. Introduction

1.1 Overview

This thesis examines acoustic and perceptual cues to distress produced by actors and real-life victims in life-threatening situations. As such, it lies at the intersection of emotional speech research and forensic speech science. Data from forensic scenarios provide a unique opportunity to study naturalistic emotions. At the same time, understanding the properties of distress productions can be of significant assistance to forensic speech scientists, if and when they are called to assess such data.

Specifically, the research presented in the following chapters explores phonetic variation both within and across two populations in a forensically-relevant context. The populations are comprised of genuine victims and actors pretending to be victims. The victims were recorded when experiencing extreme distress and were in need of emergency assistance following a violent attack. The actors were recorded re-enacting similar scenarios. Speech productions from the actors and victims are compared and contrasted using auditory-acoustic analyses in two conditions: distress (i.e. speech and vocalisations produced during a violent attack) and reference (i.e. non-distress, or baseline, speech material). In addition, a perceptual experiment is conducted using brief extracts from both actors' and victims' productions in order to examine listeners' perceptions of acted and authentic distress.

The aim of this research is to advance the process of identifying and characterising distress speech. The approach taken here is that non-speech cues are used in order to classify genuine distress productions as such, i.e. the context of the production determines the presence of genuine distress speech. Admittedly, there is no way of ascertaining whether, and to what extent, the speaker experienced distress during the speech production; however, as a first step in distress speech research, the context of a violent attack generates a plausible distress situation. The research characterises and compares phonetic properties of authentic and acted distress productions in order to investigate whether actors and victims differ in their vocal responses. Next, a perceptual study explores listeners' perceptions of distress using stimuli from the acted and authentic data. This approach advances the process of identifying distress

speech in this specific context, i.e. the victim requiring assistance from the emergency services following a violent attack, but it is hoped that further research will be able to build on the current investigation in order to refine the process of distress speech identification and thus ultimately be able to distinguish distress more broadly, e.g. from amongst other heightened emotions and/or other contexts.

1.2 Research questions

There are two main questions that guide the research in this thesis, both of which can be further divided into sub-questions, as follows:

1. To what extent can specific acoustic measures be used to identify distress speech?
    a. Are there identifiable vocal cues to distress? If so, what are they, and how can they be characterised?
    b. Are there phonetic and acoustic features associated with authentic distress responses that can be used to distinguish them from acted ones?

2. Can listeners perceive the difference between authentic and acted distress?
    a. Does listeners' accuracy improve as a function of familiarity with forensic material?
    b. Are listeners confident at distinguishing between authentic and acted distress?
    c. Do listeners' confidence levels vary as a function of familiarity with forensic material?
    d. Can listeners differentiate men's and women's distressed voices?

The first research question is addressed by conducting an acoustic study of acted and authentic distress responses. The second question is addressed by conducting a perceptual study.

1.2.1 Vocal cues to distress

Given the lack of literature on distress speech, it is not known whether there is an acoustic parameter, or combination of parameters, that can be used to identify distress speech. It seems intuitively obvious that distress brings out extreme vocalizations. Yet, although some types of extreme vocal performance have already been documented, e.g. singing (Sundberg & Rossing 1990) or shouting (Rostolland 1982; Traunmüller & Eriksson 2000), speech productions in distress, with a special emphasis on acoustic parameters, have not been quantified. An additional problem is the lack of an established frame of reference when describing distress speech. I address this by introducing a taxonomy of distress speech.

1.2.2 Differences between acted and authentic distress

Actors (and criminals attempting to deceive the authorities) occasionally need to emulate distress. Indeed, given the preponderance of acted distress in the media, most people who are not emergency services personnel, are exposed to acted distress more frequently than genuine distress. It is therefore important to determine whether there are any acoustic parameters that can help distinguish between acted and authentic distress.

1.2.3 Listeners' perceptions of acted and authentic distress

Further to the preceding research question, are listeners able to distinguish between acted and authentic distress impressionistically without access to acoustic tools? A listening experiment involving responses from lay people, forensic practitioners and police call takers will address this question.

1.2.4 Listeners' accuracy as a function of familiarity with forensic material

On the whole, many lay people would not have the occasion to listen to authentic distress from a life-threatening situation. In fact, their understanding of distress and their exemplars of distress may be limited to acted portrayals in film and TV. We can question whether their perceptions of distress are different from those of people who are familiar with and have experience of working with forensic material, such as forensic practitioners and police call takers.

1.2.5 Listeners' confidence level when distinguishing between acted and authentic distress

The literature on earwitness's ability to identify voices indicates a lack of correlation between listener accuracy and confidence scores, especially when unfamiliar voices (e.g. Yarmey (1995), (2007)) and non-modal or atypical speech such as whispered speech (Yarmey et al. 2001) are involved. Given these findings, it is of interest to investigate whether there is a similar lack of correlation between listeners' ability to distinguish real and fake productions of distress and their perceptions of their own ability. Overall, are listeners aware of their abilities?

1.2.6 Listeners' confidence level as a function of familiarity with forensic material

As an extension of the previous question, it is further worthwhile to explore whether training and experience affect listeners' confidence levels. This question is of more direct relevance to forensic practitioners, as they may be called to distinguish between acted and authentic distress as part of their practice, whereas lay people are unlikely to be asked to do so.

1.2.7 Listeners' differentiation of male and female voices in distress

Typically, listeners are able to differentiate male and female voices in a variety of conditions, such as whispered, filtered and connected speech (Lass et al. (1976); Bachorowski & Owren (1999)). These findings indicate that fundamental frequency and vowel formant characteristics are the most important acoustic correlates of speaker sex. However, if distress productions are realised with a high fundamental frequency in both male and females, will listeners be disadvantaged when attempting to attribute distress productions to a specific sex?

1.3 Background

Voice is important in our daily lives. Indeed, it has been claimed that people listen to voices more than any other sound (Belin et al. 2002: 17). When we hear a voice, we are able to infer from it a variety of information, from sociolinguistic information (e.g. regional origin, social status), to biological information (e.g. health), and emotional state. Our ability to make such inferences is not always perfect, but is usually reliable enough to facilitate our social interactions (Juslin & Scherer 2005:

65). The understanding of how we produce and perceive identifying information in the speech signal is of key importance in our everyday lives. Our ability to process vocal information may affect our appraisal of a situation and influence our response or reaction to it. Therefore, theoretical models of how we achieve this are of paramount importance to researchers interested in sociophonetic production and perception (Foulkes & Docherty 2006).

In addition, since technology plays a vital part in our day-to-day endeavours, from automated telephone banking to checking train schedules or using voice-based search on mobile devices (e.g. Siri on the iPhone/iPad), we now rely on computer systems to process vocal information for recognition and identification purposes, and also have a greater demand for natural-sounding speech synthesis (Latinus & Belin 2011). Advances in speech science research have led to the intelligibility of synthetic speech matching that of human speech, yet natural-sounding emotional expression in synthetic speech remains difficult to achieve (Schröder 2001). Applications of emotionally-expressive synthetic speech include the development of voice prosthesis systems that enable speech- and motor-impaired individuals to communicate, such as the Helpful Automatic Machine for Language and Emotional Talk (HAMLET) system (Murray et al. (1991), Murray & Arnott (1995)). In computer games, expressive emotional speech synthesis may be applied to virtual characters in order to improve the user's experience and provide a more interactive game (Gebhard et al. 2008). The development of computer-based emotional speech recognition models has applications in call centre environments and automated telephone systems as part of a decision support system to facilitate prioritising and responding quickly to agitated customers (Petrushin (1999), Morrison ete al. (2007)). Research on emotional voice production and perception is therefore not just of interest to the areas of speech science, but also psychology, medicine, engineering, and computer science.

Furthermore, the analysis of speech and sound has an increasing presence within criminal investigations. Given the widespread availability and use of mobile telephones, criminal activity is frequently audio-recorded (and sometimes also video-recorded) by victims, witnesses and even the perpetrators themselves. These

recordings may serve as evidence in court cases. Those with experience in speech and sound analysis may find themselves called as expert witnesses to provide the court with information concerning the evidence where lay knowledge would be insufficient. The increase in demand for this type of expertise has led to the recent development of Forensic Speech Science (FSS). FSS knowledge has been applied to cases involving, amongst others, speaker comparison, speaker profiling, disputed utterance resolution, voice parades, enhancement, authentication and transcription.[1] The research presented in the following chapters concerns an area of research situated firmly within the scope of FSS and emotional speech, being a forensic phonetic study of the vocal responses of individuals in distress as a result of a violent attack.

1.4 Motivation

The research questions have been motivated by the following considerations:

1.4.1 Theoretical relevance

Samples of extreme emotional speech are difficult to obtain. Firstly, for ethical reasons, it is not possible to induce extreme emotion in an experimental setting. Secondly, where speech data does exist, e.g. produced in a real-life situation, it may not be available publicly due to data protection and privacy issues, as well as potential legal constraints. Therefore, analysis of authentic speech data from recordings used in previous criminal investigations, specifically those cases that are no longer part of a judicial process, represents a unique opportunity to explore the effects of extreme emotion on the human voice. Research in this area assists speech scientists to categorise and identify cues to emotional speech, to explore the limits of possible human vocalisations, and to develop understanding of vocal tract function in extreme conditions. Furthermore, data from forensic recordings often present major analytic difficulties, e.g. they may be brief in duration, or of inferior quality, or contain emotive speech, and consequently acoustic components can differ from those depicted in traditional phonetic studies in laboratory settings.

[1] For further details on the breadth of applications of FSS, readers are directed to Foulkes & French (2012), French & Stevens, (2013), and Jessen (2008). An accessible historical perspective on key issues and applications of FSS is also presented in Eriksson (2005).

1.4.2 Practical FSS relevance

Forensic speech scientists are often questioned by instructing parties, e.g. police officers or solicitors, as to whether audio recordings allegedly representing violent attacks are genuine or hoaxes, and to what extent speech occurring in forensic material reflects real distress. In the case of *State of Florida -v- George Zimmerman*, in which the defendant was accused of murdering Trayvon Martin in February 2012, the question arose whether a specific category of emotional speech (screams) could be attributed to an individual based on his/her 'normal' (i.e. reference) speech material. More recently, during the earlier stages of *The State vs Oscar Pistorius* trial, in which Oscar Pistorius was accused of murdering Reeva Steenkamp in February 2013, a prosecution witnesses (a neighbour) was questioned about the screams she claimed to have heard the night of the murder, specifically whether she thought she could tell whether it was a woman or man screaming.

Forensic practitioners who are members of the International Association of Forensic Phonetics and Acoustics (IAFPA) are currently prohibited by the IAFPA code of practice (IAFPA, 2004: clause 9) from making statements regarding the psychological states or sincerity of speakers in forensic recordings, as very little is known about how these states are manifested in vocal performances. In the Zimmerman trial, conflicting opinions were presented by the prosecution and defence expert witnesses, resulting in the judge ruling that the testimonies from the prosecution be excluded from the trial (court order document excluding evidence from Mr. Owen and Dr. Reich, State -v- Zimmerman). The Zimmerman trial demonstrates a lack of consensus in professional opinion in this area, and clearly highlights the consequences of such a lack in real-world trial situations and the need for further study. In the Pistorius trial, the attribution of the screams by the witness was called into question; the prosecution argued that the screams were produced by from the victim, supported by the testimony of the witness, whereas the defence claimed that the screams the witness heard were produced by the defendant. Research on emotional speech in forensic situations would represent the first step towards results that might ultimately be used to substantiate forensic expert opinions in this area.

1.5 Thesis outline

The thesis is divided into nine chapters, including this introductory section. Preceding the description of the structure itself is an overview of the study, an account of the research questions, and the motivations behind them.

Chapter 2 contains a review of the academic literature on emotional speech produced in contexts comparable to forensic scenarios. It critically summarises studies concerned with production and perception of emotional speech, particularly those concerned with fear and stress.

Chapter 3 illustrates the methodology adopted. It presents the merits of combining authentic and acted data for the purposes of the current investigation, introduces the two datasets, and describes the data collection, analysis techniques and parameters under investigation.

Chapter 4 describes a pilot study that was used to validate an aspect of the methodology (the distress taxonomy) before it was fully adopted. It also reports on findings from a small-scale perceptual experiment that was conducted as part of the validation process.

Chapter 5 presents the results of the acoustic analyses. Findings for each parameter are described. First, findings from the authentic data are presented, followed by those from the acted data, and finally the findings are compared across the two datasets.

Chapter 6 expounds the methodology used to conduct a perceptual experiment where perceptual cues to distress are explored using both datasets as stimuli. It tests whether different groups of listeners can distinguish acted distress from real distress impressionistically.

Chapter 7 presents the results of the perceptual test by examining listeners' responses to the stimuli, as well as the stimuli themselves.

Chapter 8 considers the implications of the results reported in chapters 5 and 7 for the field of forensic speech science and emotional speech research generally, in terms of the research questions presented at the beginning of this thesis.

Chapter 9 summarises the research and reflects on its contribution to the field. It highlights areas for future research which may further develop our knowledge of the production and perception of distress speech.

2. Review of Literature

This chapter summarises the principal literature concerned with emotional speech, with specific reference to speech produced by individuals experiencing stress and fear. The first part of the chapter presents a historical perspective on emotional speech research. The second section presents a brief description of the physiological changes involved in emotional speech. The third part discusses challenges in emotional speech research in terms of defining and conceptualising emotion and stress. The fourth section reviews previous empirical production and perception studies of emotional speech according to the methodology used. It identifies acoustic correlates of fear and stress, and it describes emotion recognition accuracy rates. The fifth part discusses an analysis of screamed productions. A table is provided summarising the literature in the sixth section.

2.1 A historical overview of emotion and voice

Scherer (2003) highlights that the systematic study of emotion in speech and its effect on the listener can be traced back to classical Greek and Roman rhetoric grammars (e.g. *De Oratore* by Cicero, and *Institutio Oratoria* by Quintilian), who drew a focus on improving readers' oratorical skills. The focus on emotion in rhetorical speech remained a feature of western philosophy for the centuries that followed (see Kennedy (1972)). An early scientific approach to emotional speech can be found in the work of evolutionary biologists in the nineteenth century, such as Darwin. Indeed, many questions about vocal expression of emotion nowadays have their origins in the ideas and notions first put forward by Darwin in his seminal work *The Expression of the Emotions in Man and Animals* (Darwin 1872). Although motivated mainly as a treatise to support his evolutionary theories (and not purely by an interest in the communication of emotion), the research was the first scientific study of emotional expression, and the first to seek people's perceptions of emotions portrayed in facial expressions as a way to explore the meaning of expression (Hess & Thibault 2009: 126).

Empirical research on vocal expression began in earnest during the early twentieth century following the development of sound-recording technology. Studies from this

period, such as those by Scripture (1921) and Skinner (1935), investigated emotional vocal production with the principal aim of assisting diagnoses of psychiatric disorders (Juslin & Scherer 2005: 67). The emergence of the radio and telephone resulted in further interest in vocal expression (Scherer 2003: 228). By the middle of the twentieth century, the focus of emotional voice studies had shifted to analysing changes in the voice as a means to measure and monitor the emotional state of aviation personnel such as astronauts (Simonov & Frolov 1977) and pilots (Williams & Stevens (1969), Kuroda et al. (1976)), in line with advances and international interest in aviation technology at the time.

In the latter half of the twentieth century, systematic investigation of emotional speech further developed, albeit in a disjointed fashion across disciplines (Scherer 2003). Psychologists studied emotional expression through different modalities, e.g. Ekman (1971; 1992), and Izard (1971; 1977). Linguists investigated the importance of pragmatics in emotional interactions (Caffi & Janney 1994). Advances in spectral analysis and speech recording equipment provided engineers, phoneticians and computer scientists with an opportunity to examine vocal expression of emotion using increasingly sophisticated technology, e.g. Williams & Stevens (1969, 1972) (1969; 1972), Klasmeyer & Sendlmeier (1997), and Burkhardt & Sendlmeier (2000). Furthermore, thanks to developments in technology and a greater demand for natural-sounding speech synthesis, speech scientists and engineers also began to devote more research to emotional expression, developing new disciplines such as ‘affective computing’ (Picard 1997). The study of speech under stress, which is related to emotional expression, also gained momentum, resulting in the European Speech Communication Association (ESCA) and North Atlantic Treaty Organisation (NATO) interdisciplinary workshop in Lisbon, Portugal, which centred on definitions and models of stress (Moore & Trancoso 1995), and a special issue of speech under stress in *Speech Communication* (1996). More recently, forensic speech scientists have investigated the acoustic effects of emotional voice and speech under stress in forensic situations, e.g. Jessen (1997), Meinerz (2008), Kirchhübel & Howard (2013).

By the start of the twenty-first century, however, it was recognised that emotional speech research would be more likely to yield advances if addressed using interdisciplinary approaches, e.g. Davidson et al. (2003). Collaborations between academics who come from backgrounds such as phonetics, speech processing and psychology are increasingly taking place, as demonstrated not only by the aforementioned ESCA-NATO workshop on speech under stress, but also the first International Speech Communication Association (ISCA) Speech and Emotion workshop held at Newcastle, Northern Ireland, in 2000 (Douglas-Cowie et al. 2003), leading to the special issue of speech and emotion in *Speech Communication* (2003). These days, emotional speech papers regularly appear in a variety of conferences and journals. Following these trends in both forensic and interdisciplinary approaches, the present study incorporates a production and perception study of a specific type of emotional speech, that of distress as experienced by an individual experiencing a violent attack.

2.2 A physiological overview of emotion and voice

Psychologists specialising in emotion generally accept that emotions result in a variety of adaptive responses by the nervous system. These, in turn, lead to changes in the production of the speech signal. A summary of this research can be found in Johnstone and Scherer (2000: 222). Consequently, when faced with a stressful situation, e.g. a threat, changes to physiology are produced as a survival mechanism (Kirchhübel et al. 2011: 77, citing Jessen, 2006:23). Scherer (1979; 1981) highlights that three changes in particular - increases in respiration rate and muscle tension, and a decrease in saliva production - will lead to changes in speech production.[2]

Firstly, a more rapid respiration rate produces an increase in sub-glottal pressure which may manifest itself acoustically as a shift of energy to higher frequencies and/or an increase in the amplitude of vocal fold vibration. The increase in sub-glottal pressure (and subsequent increase in supra-glottal pressure and airflow) may

[2] Other physiological changes include a release of adrenaline, sharper cognitive and sensory skills, pupil dilation, and increases in cardiovascular activity and perspiration (Kirchhübel et al., 2011:77, citing Jessen, 2006:20-23).

lead to greater turbulence and friction, as well as faster vocal vibration due to a more intense action of the Bernoulli Effect (Jessen et al. 2007).

Secondly, an increase in laryngeal muscle tension may result in an increase in F0 due to the increased tension of the vocal folds. Muscles in the jaw, lips and tongue may also be affected by increased muscle tension, leading to articulatory undershoot or overshoot of target sounds. As summarised by Kirchhübel et al. (2011: 81), target undershoot may be observed if the speaker is unable to reach the consonantal/vocalic targets due to the increased muscle tension. In vowel production, for example, this may take the form of a contraction of the vowel space. Target overshoot may occur when a tensed tongue produces a faster and more forceful movement for the target sound, while the speaker is unable to exhibit much control over the timing of the gestures. Furthermore, a lack of synchronization between the laryngeal muscles used in phonation and respiration may also lead to voicing irregularities (Scherer 1979).

Thirdly, a decrease in saliva production has the effect of increasing energy of harmonics due to the vocal tract having drier surfaces (Scherer 1986). This may result in a narrow bandwidth of the formants, and voices possessing this characteristic are typically labelled impressionistically as, amongst others, 'metallic', 'piercing' and/or 'strident' (Scherer 1986: 152).

Although these three physical changes may result in predictable acoustic outputs, variation across speakers' productions can be expected due to variation in an individual's perception and evaluation of the threat, i.e. psychological factors mediate their physiological responses, thus resulting in an individualisation of the response (Hollien 1980; Kirchhübel et al. 2011).

The process by which emotions (in a general sense) are negotiated by way of a subjective personal evaluation of an event, e.g. a threat, has been another area in which a substantial amount of psychological research has been conducted. A full overview of this research is beyond the scope of the current discussion, but it is worth considering some of the themes that have arisen therein. A prominent and popular theory that attempts to describe emotional responses along these lines is

Appraisal Theory, championed by Lazarus and colleagues, e.g. Lazarus (1966; 1991) and Lazarus et al. (1970). In Appraisal Theory, subjective evaluation plays a significant role in stress/emotional reactions, thus accounting for variation in an individual's response. Moreover, Lazarus (1966) recognises that the dynamic nature of appraisal can result in a reappraisal of the situation based on new information or re-evaluation, thus allowing not only inter-individual variation, but also intra-individual variation in responses across different circumstances and even within a particular event as it develops. Thus, the very nature of the relationship between the psychological reaction and the physiological approach is predicted to be dynamic and multi-layered.

In addition to the relationship between physiological and psychological responses to emotion (or in the above case, stress), a relationship between psychophysical responses and socio-cultural responses in vocal expression of emotion needs to be acknowledged. Johnstone and Scherer (2000: 223) state that the human communication system as it is today has evolved to take advantage of two distinct systems: a non-verbal vocal call system that is traditionally thought to signal emotion, and a verbal communication system which allows us to speak and write. Our original vocal call system is often likened to non-verbal call systems in other species, such as grunts and alarm calls in vervet monkeys (Seyfarth & Cheney 1986). It is unclear how one system may constrain the other, i.e. how the evolution of our speech system has been constrained by emotion signalling, or how emotional expression has been constrained by human speech, though evidence shows the two systems can function independently of each other. For example, Scherer et al. (1984) conducted a perceptual experiment to test whether non-verbal cues function independently of verbal communication when judging emotion. Participants were asked to judge both audio and written stimuli for 'tone' of the extract using a prepared list of adjectives such as 'polite', 'insecure', and 'aggressive'. Some audio extracts were filtered, rendering them unintelligible, so that listeners would base their judgments on purely non-verbal information. It was found that both verbal and non-verbal information contributed to perceiving communication of emotion, but that it was also possible for participants to correctly gauge emotion from just the

filtered audio samples, suggesting that non-verbal cues function both parallel to, and independently of, verbal communication.

Reflexive physiological processes, such as changes in respiration and muscle tension as responses to a stimulus (e.g. a threat), are typically associated with the non-verbal system. These are known as “push effects” since they can act as pushing emotional expression in a certain direction (Johnstone & Scherer 2000). For example, increased muscle tension and respiration rate due to fear can lead to an increase in F0, leading to a higher-pitched vocal production.

In contrast, “pull effects” - external factors such as social norms, physical conditions or listener expectations - pull expressions in a different direction (Scherer et al. 1980). Push effects are involuntary processes that directly influence the vocal parameters to shape a vocal production, whereas pull effects are externally-based and shape the vocal production by orienting to a specific acoustic target (Scherer 1988). Both push and pull effects influence vocal expression of emotion.

This distinction can be clarified using an example provided by Scherer (1988: 82). If an individual were preparing oysters at home and was confronted with a slithering worm that emerges upon opening a shell, the individual may respond with a high-pitched, “Eee!” This type of production acts in the same way as animal calls in that it communicates disgust and warns others to be wary with their own unopened oysters. It is an example of a production influenced by a push effect, since the physiological response to the shock and disgust may lead to increased muscle tension and in turn an increased F0. In contrast, if the same individual observes someone else eating oysters containing worms, s/he may produce the response, “ Yuck!” In this case, the production still expresses disgust, though in a culturally-specific way. The production is not influenced by physiological responses; instead, it is governed by social conventions.

2.3 Challenges of emotional speech research

Researchers investigating vocal expression of emotion face a variety of challenges. There are often studies with a high degree of overlap due to a lack of uniformity in

defining and conceptualising the area of research under investigation. As such, methodological approaches to investigating emotional expression also vary widely. As a consequence of these issues, it is often challenging to compare studies directly.

2.3.1 Defining the area of research

A major concern for emotional expression is that terminology across the disciplines is inconsistent and often loosely defined. As observed by Scherer (2005: 696):

> *The concept of ‘‘emotion’’ presents a particularly thorny problem. Even though the term is used very frequently, to the point of being extremely fashionable these days, the question, ‘‘what is an emotion?’’ rarely generates the same answer from different individuals, scientists or laymen alike.*

Scherer (2000: 138) comments that divergences occurring across definitions of emotional speech typically arise from issues such as:

1. which changes across different modalities are important to emotion, and to what extent these changes are the product of the emotion;
2. whether it should be assumed that emotions are triggered by internal/external stimuli (or events) rather than viewing them as states which are relatively stable over time.

A lack of agreement in terminology means that there is often a difference in focus of studies across disciplines. Juslin and Scherer (2005: 67) comment that linguists have been criticised for failing to embrace developments in psychology, and psychologists have been criticised for neglecting language and interaction. In a similar vein, research which centres on emotion may also be applicable to stress, and research on stress may also be applicable to emotion. Juslin and Scherer (2005: 70) acknowledge that much of the work investigating emotional expression falls into two main classes: those that investigate speech under a particular type of stress, e.g. workload, time pressure, cognitive stress, physical stress, etc.; and those that investigate speech through emotion, e.g. anger, happiness, fear, sadness, etc. Intuitively, there is considerable (yet hitherto poorly-defined) overlap between the two. Hollien (1980: 48) asserts that stress relates to psychological states, often referred to as emotions,

yet it needs to be recognised that not all emotions accompany stress (e.g. joy and happiness). Until recently, speech under stress was subject to more research in speech science than was speech associated with specific emotions (Johnstone 2001: 17).

Juslin and Scherer (2005), acknowledging that a wide array of terms such as 'emotion', 'affect', 'mood', 'stress', etc. are treated synonymously, thus causing confusion within the vocal communication field, propose that *affect* be the umbrella term given to describe the variety of states such as emotion and stress, mood and interpersonal stance, and preferences and attitudes, that have 'affected' the individual. These affective states can then be further distinguished along the dimensions of intensity of response, duration of response, synchronisation (the degree of co-ordination between the individual's different biological systems), event focus (whether an affective state is triggered by a specific event), appraisal elicitation (the degree to which personal evaluation impacts the affective state), rapidity of change, and behaviour impact (strength of change to the individual's behaviour). Figure 2-1, adapted from Juslin and Scherer (2005) and Scherer (2000), illustrates these distinctions for the affective states of emotion and stress, since these are the two most relevant to the current investigation.

Figure 2-1: The differentiation of stress and emotion based on feature dimensions proposed by Scherer (2000), adapted from Scherer (2000) and Juslin & Scherer (2005).

Affect

Affective state:	**Emotion**	**Stress**
Dimension:		
Intensity	++	++
Duration	.	+
Synchronisation	++	++
Event focus	++	+
Appraisal elicitation	++	++
Rapidity of change	++	+
Behaviour impact	++	++

++ = medium/high
+ = low/medium
. = absent/low

Juslin and Scherer (2005) group emotion and stress together due to their similarities, since both are considered to be powerful, short, intense responses to an event of high

importance to the individual, and involve a lot of synchronised changes in the physiological and/or psychological systems of the individual. Both affective states also dramatically change the individual's behaviour, though it is recognised that in both cases there may be changeability of the response following re-appraisal of the event. Both affective states have similar dimensions, but stress can be differentiated from emotion by its longer duration, the low likelihood of it being triggered by a one-off event, and its tendency to change state quickly. In contrast, mood and interpersonal stance are weak, rarely synchronised responses which last a long time. They are rarely triggered by an event that requires appraisal. Moods may be triggered by subconscious factors, such as fatigue or hormones, whereas interpersonal stance is the way in which two or more individuals interact, e.g. in a warm and friendly way versus a cold, distant way, and may be present either intentionally or subconsciously. Lastly, preferences and attitudes generally consist of low-intensity, low-synchronised responses with little change to the individual's behaviour responses, though little research has been conducted on cues to these types of affective states.

Terminological differences are also ubiquitous in research on speech when under stress. Hollien (1980), Murray et al. (1996), and Kirchhübel et al. (2011), all discuss a similar problem in studies of speech when under stress. There are many interpretations of stress, including, amongst others, mental strain, emphasis, and force, though most disciplines employ the term 'stress' to indicate either psychological- or physiological-based tension and strain, or a combination thereof. However, even when practitioners in different fields agree on this interpretation, a single definition that is able to satisfy researchers in all disciplines is lacking. As stated by Cox (1978:1):

> *The concept of stress is elusive because it is poorly defined. There is no agreed definition in existence. It is a concept which is familiar to both layman and professional alike; it is understood by all when used in a general context but by very few when a more precise account is required, and this seems to be the central problem.*

Murray et al. (1996) comment that researchers at the ESCA-NATA workshop on speech under stress were unable to arrive at a single definition of stress that would satisfy all those present. Instead, six definitions were discussed and were all considered 'correct'. One of the main areas in which researchers failed to reach agreement was whether stress is considered to be a cause or an effect (or neither). It was found that defining stress via stressors, i.e. the stimuli that produce stressed speech, was a more fruitful endeavour and so a taxonomy of stress incorporating different types of stressors was introduced. Stressors are classified as either 'physical', 'chemical', 'physiological' or 'psychological', and can be further divided into sub-categories based on the stage at which the speech production chain is affected most by the stressor. Zero-order stressors are those that result in external physical changes or obstructions to the vocal apparatus due to, for example, vibrations affecting the articulators. First-order stressors are those from unconscious physical changes in the body such as changes in muscle tension or breathing rate. Second-order stressors are conscious physiological changes addressing physical constraints of the environment such as an increase in vocal effort due to lots of background noise. Third-order stressors are those that incorporate the individual's psychological state into the speech production, which may result in an unconscious physical change. For example, an individual may shout out at his/her interlocutor in order to be heard because there is a lot of background noise (a second-order stressor) or an individual may shout at his/her interlocutor in a non-noisy environment because s/he is angry or upset (a third-order stressor). Table 2-1 and Table 2-2, adapted from Murray et al. (1996), and Kirchhübel et al. (2011), display the taxonomy of stressors and the taxonomy of stressor order, respectively.

Table 2-1: The taxonomy of stressors, as adopted by the ESCA-NATO workshop members, adapted from Murray et al. (1996) and Kirchhübel et al. (2011).

Classification	Stressor
Physical	noise, physical exercise, vibrations
Chemical	fatigue, alcohol, drugs
Physiological	illness, disease, injuries
Psychological	emotion, depression, workload

Table 2-2: The taxonomy of stressor order, as adopted by the ESCA-NATO workshop members, adapted from Murray et al. (1996) and Kirchhübel et al. (2011).

Stressor Order	Stressor order description
0^{th}	external, physical changes/obstructions affecting the vocal apparatus, e.g. oxygen masks (Fecher & Watt 2011)
1^{st}	internal, unconscious, physical changes caused by, e.g. alcohol (Chin & Pisoni 1997)
2^{nd}	internal, conscious, physical changes caused by response to environment, e.g. Lombard speech (Junqua 1996)
3^{rd}	internal, physical changes caused by psychological state, e.g. pilots experiencing aviation problems while flying (Williams & Stevens 1969)

Overall, the relationship between physiological and psychological states is less straightforward than that between physiological states and voice. Not only is there more inter-subject variation in how psychological states are manifested, but also the effects of psychological changes on voice are indirect and mediated by the speaker's physiology. Therefore, while hypothetically it may be possible to fully predict the effects of a particular physiological or environmental condition on the voice, and it may be possible to infer the speaker's physiological conditions from the acoustic signal, it is considerably more challenging to predict the vocal effects of a psychological state.

Thus far, this chapter has reviewed historical, physiological, and terminological difficulties, taking a rather broad view of emotion speech research, despite the fact that the current investigation focuses on distress, a specific type of affective response produced in forensic contexts such as violent attacks. The primary reason for considering such a broad area is the fact that very little literature exists that focuses purely on distress speech. Instead, in psychology or linguistics research, if distress speech features as part of an investigation, it is usually treated alongside other emotional states (e.g. Chung (2000)), typically those that comprise the 'Big 6', i.e. happiness, sadness, fear, disgust, anger, and surprise (Cornelius 1996), rather than as an area of research interest in its own right.

In the clinical domain, distress speech research is often associated with long-term physical and psychological disorders, such as analysing the speech of social phobics before and after a social phobia treatment (Laukka et al. 2008). A related issue in

clinical research is the relationship between psychological distress as a cause of voice disorders, where the patient's difficulties arise not from a physical problem but because the patient subconsciously wishes to avoid communication, and psychological distress as a result of a voice disorder, where patients find their difficulties frustrating. In both cases, speech will manifest signs of both the disorder and the distress, but different treatments may be called for (Seifert & Kollbrunner 2005).

In computer science research, automatic speech and non-speech sound classification and speech recognition systems have been trained to distinguish between normal, every-day sounds and abnormal sounds (those originating from a distress situation) from around the home in order to act as audio sensors as part of a remote monitoring system for the elderly living at home (Vacher et al. 2006).

The above examples show that research on distress speech is fruitful, yet in different, and not always connected, ways. Common to all of the above studies is that distress is not explicitly defined. Instead, distress is assumed to be a concept that is understood by all parties and is qualified by a specific investigation/situation. There is currently no research on distress speech for forensic purposes, despite the fact that the very nature of criminal activity, e.g. situations involving assault and violence, will often result in distress speech produced by victims and witnesses. Similarly, there is no research being carried out directly comparing the distress speech of real victims and the simulated distress speech of actors.

2.3.1.1 Defining distress

As demonstrated in the previous section, the definition and use of the terms "emotion", "stress" and "distress" are not uncontroversial. For the purposes of this research, the term 'distress' is used specifically in order to refer to the affective states such as fear and psychological stress that arise from violent attacks recorded in authentic forensic material. The definition employed by Hicks (1979: 12), and later adopted by Hollien (1980), has been used as a basis for this research: "*stress . . . is a psychological state that is a response to a perceived threat and it will be accompanied by specific emotions*". In order to be of relevance in a forensic context,

then, distress will be viewed as a psychological state that is a response to a perceived threat *as a result of a violent and/or life-threatening attack* and will be accompanied by specific emotions. In terms of the ESCA-NATO taxonomy of stressors, distress of this type would be classified as a 'psychological' and third-order stressor, though in some cases, it will be accompanied by additional stressors of a different nature (for example, injuries to the chest, neck or head may also lead to a 'physical' effect on vocalisation above and beyond the psychologically-based effect of distress itself).

Of the most commonly used 'Big 6' emotions (Cornelius, 1996), 'fear' would perhaps be considered the most appropriate to forensic scenarios. Both psychological and physiological stress may also play a role in distress speech, and so studies involving different types of psychological and physiological stress are also presented. Fear and stress are of course not synonymous terms, and it is important to distinguish between the two. Instead, I report on those studies that are applicable to the topic under investigation in this thesis, that of speech of individuals in distress.

2.3.2 Conceptualising emotion

As suggested by the lack of standardisation with regard to the terminology relating to affect and emotion, the conceptualisation of these ideas is equally lacking in consensus. Three principal models which focus upon how emotion can be conceptualised dominate the field of emotional expression.[3]

2.3.2.1 Discrete emotion models

In these models of emotion, it is proposed that there are a limited number of fundamental emotions such as 'fear', 'anger', 'joy', 'sadness' etc. that have developed as adaptive responses during the course of evolution. Proponents generally distinguish between 7 and 14 emotions (Scherer 2000: 147) which are based on the major emotion terms described in Darwin (1872). In Darwin's seminal study, he proposed that emotions are discrete states, and for each one he described its evolution, functionality and universality across both humans and animals.

[3] See Scherer (2000) and Juslin & Scherer (2005) for a more detailed overview of this topic.

According to this approach, emotions are considered to be associated with specific eliciting conditions and specific vocal and facial responses. Ekman (1971; 1992) and Izard (1971; 1977), in particular, have popularised the discrete emotion model, especially in the field of psychology, by extending the theory and investigating facial expression of emotion empirically. Many present-day emotion studies have been influenced by their work and many of the emotion labels are in frequent use, especially the 'Big 6' emotions (happiness, sadness, fear, disgust, anger, and surprise (Cornelius 1996)), which are the most commonly found across studies. With respect to vocal expression of emotion, many studies have investigated the use of specific vocal profiles for each emotion, both in psychology research (Juslin & Scherer 2005), and in the field of speech technology (Cowie et al. 2001).

In recognition of the fact that there may be different kinds and gradients of the same emotion, e.g. hot, uncontrolled anger vs. cold, controlled anger, and in order to take into account a lack of uniformity in vocal profiles of the alleged same emotion, the discrete emotion model has been further extended to include the concept of 'emotion families' (e.g. Ekman (1992)), whereby each category is taken to represent a grouping of related emotional states. Emotion families allow the models to retain their discrete nature at the level of broader emotion categorisation, while also accounting for variability within the same emotion category.

2.3.2.2 Componential models of emotion

At the heart of the componential model is appraisal theory. This model of emotion assumes that emotions are elicited through the individual's cognitive evaluation, either consciously or subconsciously, of a stimulus or event, such as a threat, which determines different physiological and expressive responses. Cognition is considered an antecedent of emotion. This approach explicitly highlights the link between the emotion elicitation and stimulus appraisal, and proposes that the distinction between emotions can be made with reference to these links. In contrast, in discrete emotion models the difference between emotions is attributed to an evolutionary response to a stimulus.

The componential model of Lazarus (1991) is similar to the discrete emotion model in that there are finite number of appraisal themes which trigger a finite number of

fundamental emotions, yet there is more emphasis on the process of elicitation and appraisal. It can be considered a restrictive componential model, unlike the component process model proposed by Scherer (1984). The assumption behind Scherer's component process model is that there are as many emotions as there are combinations of appraisal responses. Using the model, Scherer (1986) posits a list of predicted vocal and acoustic changes based on physiological responses to particular appraisals.

Lazarus and Scherer represent the two extremes of the componential model continuum. Most proponents of componential models represent intermediate points between these extremes but the general consensus is that use of limited set of fundamental emotions is not an accurate representation of emotional expression. It is, however, useful to consider super-ordinate emotions that form families or prototypes of emotions.

2.3.2.3 Dimensional models of emotion

Models of emotion that map affective states onto a specific dimension (or combination of dimensions) are known as dimensional models. One-dimensional models are those in which one dimension, usually either valence (e.g. pleasant/unpleasant) or activation (e.g. active/passive, sometimes referred to as 'arousal'), is sufficient to distinguish between the emotions. Multi-dimensional models are those that operate on two or more dimensions, typically the two described above, and if a third dimension is added, it concerns a 'power' dimension. Studies investigating the activation dimension of emotion generally provide consistent acoustic data (see e.g. Laukka et al. (2005), Schröder et al. (2001)), yet the dimensions of valence and power generate inconclusive data (though note that power is still little studied). The advantage of a multidimensional approach is that it provides a visual representation of similarities and differences between emotions.

Proponents of the dimensional model of emotion believe that the distinction between affective states can be reflected through changes in a broad physiological system, unlike the discrete emotion model, which supports the idea that emotions are the result of discrete emotion-specific physiology with universal (and evolutionary) antecedents. Cognition is viewed as a consequence of emotion by followers of the

dimension model, whereas componential model supporters view cognition as an antecedent of emotion.

2.3.3 Conceptualising stress and distress

In contrast with the study of emotion, research into distress has seen considerably less debate. Although a wide variety of definitions can be used to describe the different categories of stress, the overall approach has been consistent, as discussed above (§2.3.1). Essentially, models of stress all take a behavioural approach, employing a taxonomy in which the stimuli and their effects on the vocal tract, be they physiological or psychological, are reported.

It is worth noting that as of the time of writing, there have been no studies offering a conceptual model of any type for distress, as there simply has not been a broad enough empirical base among the few existing studies to allow for one. The present study, in expanding this base, will provide a step forward toward such a model, but as it focussed only on one type of data (violent attacks resulting in life-threatening situations), I cannot offer a general categorisation. Instead, in the chapters that follow, I will follow the stress literature by offering a taxonomy, but it will attempt to map acoustic cues to level of distress within this single data type.

2.4 The effects of emotion and stress on speech production and speech perception

This section presents a summary of the literature investigating how emotion and stress affect speech production and perception. This literature is divided into different types depending on the methodology used to procure the emotional speech.

Just as there is a lack of consensus in defining and/or conceptualising affective speech terms, there is a lack of generally accepted and standardised approaches and techniques from researchers across the various disciplines in which investigators conduct research. Rather, a host of literature describing various paradigms is available, primarily from the fields of psychology and biology (Cornelius 1996). A range of tools is therefore at the emotional speech researcher's disposal. No one

method has yet resulted in a fully integrated and multidisciplinary approach, as there are advantages and disadvantages to each one (Cowie & Cornelius 2003: 5).

Emotional speech studies traditionally employ one or more of the following three main paradigms:

1. acted portrayals of emotion
2. experimentally-induced emotion
3. genuine emotion

The majority of emotional speech studies have focussed on collecting and analysing acted emotional datasets and laboratory-induced emotions.

2.4.1 Acted portrayals of emotion

For acted datasets, actors are typically asked to portray two or more emotions, typically those from the 'Big 6', i.e. happiness, sadness, fear, disgust, anger, and surprise (Cornelius 1996). Portrayals are then examined acoustically, and comparisons drawn across the specified emotions in order to see which vocal correlates can be used to distinguish one emotion from another. These studies are often extended by implementing a perceptual test using stimuli based on the actors' portrayals. Listeners are asked to try to recognise the perceived emotion by indicating (or rating) on a response sheet the emotion they perceived. As the choice typically involves a range of 5-8 emotions (depending on the specific study), chance performance would be reflected by an accuracy rate of 12%-20%. In practice, accuracy rates of emotion recognition average 60% on the whole and are, therefore, much greater than chance (Scherer 1989). Of equal interest are the misidentifications of the portrayed emotion, since confusions can assist researchers in identifying the similarity or proximity of emotion categories (Johnstone & Scherer 2000).

Vocal responses and accounts of emotive events have in some studies formed the basis of acted portrayals of emotion. For example, Williams & Stevens (1972) recorded and analysed a professional actor imitating a radio newscaster's report of the Hindenburg air disaster (the crash of a zeppelin in 1937). This recording

consisted of a male newscaster announcing the arrival of the Hindenburg when the zeppelin suddenly burst into flames prior to docking. The newscaster continued to broadcast his report (albeit with short breaks), and so his voice was captured before, during and after the disaster. The F0 mean, range and contour of the actor and the original newscaster were compared, and similarities were observed in the form of increases in mean and range post-event by both speakers, and 'irregular' and 'atypical' F0 contours (Williams & Stevens 1972: 147-48), thus allowing them to conclude that the use of actors was justified.

The comparison of acted and authentic Hindenburg report recordings was part of a larger study involving acted emotion. Williams & Stevens employed three professional male actors and a director, all of whom had been members of the Actors Studio in New York, to enact a play that had been specially constructed to incorporate different emotional situations for the three characters that contained some identical speech material across characters and situations. The play was designed to elicit the emotions of anger, sorrow, and fear and compare the same phrases in 'neutral' speech. The parameters they investigated were F0 mean, range and contour, formant frequencies and articulation rate. They found no clear acoustic correlates of fear, but they did report that the actors' average F0 was in line with their anger and neutral speech, and that their maximum F0 was often much higher than neutral speech. When observed spectrographically, they found that the high F0 peaks present in fear speech often "had unusual shapes (irregular bumps or discontinuities)" (Williams & Stevens 1972: 1249), as well as voicing irregularities. The duration of fear speech utterances was also longer than that of anger and neutral speech utterances.

Scherer et al. (1991) employed four German professional radio actors (two male and two female) to read out sentences based on a short realistic scenario that was designed to elicit a specific emotion. For example, in order to elicit sadness, the actors were asked to read a sentence explaining the scenario that they had to give up the family pet because they were in the process of moving to a new property in which pets were not permitted. The emotions investigated were anger, sadness, joy, fear, and disgust. The recordings then acted as stimuli in a series of listening tests.

Although listeners were unable to successfully recognise ‘disgust’, the other emotions received an accuracy recognition rate of over 62% across the series of tests. Portrayals of emotion which were consistently and accurately recognised by listeners were then subjected to acoustic analyses. ‘Fear’, when compared with ‘neutral’ expression, was characterised by significant increases in articulation, intensity mean and variance, and F0 mean and variance.

Similarly, Banse & Scherer (1996) recorded twelve German professional stage actors (six male, six female) portraying fourteen different emotions, two of which are described as belonging to the ‘fear family’: ‘panic fear’ and ‘anxiety’. The authors differentiate between the two but do not state if their label of ‘panic fear’ should be treated as synonymous with other emotional portrayals of ‘fear’ across studies. The actors were recorded reading out the same standard sentences for each emotion after having read and imagined descriptions of antecedent scenarios designed to elicit the emotions under investigation. The scenarios were drawn from a large corpus of cross-cultural studies investigating emotional experiences (Scherer et al. 1986). The recognisability of emotional portrayals was then rated by advanced acting students. The two most recognisable portrayals from two different actors were chosen to represent each emotion in a listening experiment (224 portrayals), which also included four ‘next best’ portrayals of each emotion (resulting in 280 portrayals in total). Acoustic analyses were then performed on the original highest-rated 224 portrayals.

The mean recognition rate over all 14 emotions was 48% (chance being 7%), though there was much variation in accuracy rates across emotions. Panic fear had a low recognition rate of 36%. The authors explained that this was due to confusion with ‘anxiety’, and demonstrated that if the ‘fear family’ was collapsed as one emotion, the recognition rate would increase to 63%. An acoustic profile comprising of a high mean F0, high mean energy, and an increase in speech rate was put forward for ‘panic fear’.

Bonebright et al. (1996) were interested in the perception of emotion, with a focus on gender differences. Unlike the other studies described in this section, they made no attempt to analyse the acoustics of their acted samples. They used 6 trained theatre students (3 male and 3 female), who were each instructed to read two stories in fearful, angry, happy, sad and neutral ways – twice for each story/emotion combination, resulting in a total of 120 stimuli passages. Listeners were divided into two categories, which performed different tasks: judges who were asked to name the emotion which was present in each passage, and raters who were told which emotion was being attempted in each passage and had to score the effectiveness of the portrayal on a numerical scale. They found that female judges had higher identification rates for fear, sadness and happiness compared to male judges, while male judges were better at identifying anger. They also found that male actors' portrayal of anger and fear were identified better and rated higher than female performances of those two emotions. The authors suggest that these differences could be the outcome of a socialisation process that begins in childhood.

Leinonen et al. (1997) employed non-professional actors, eight male and eight female, to produce the name 'Sarah' in Finnish [saara] in ten different emotion portrayals (including 'frightened') based on ten different frame stories. For example, to portray [saara] in 'frightened' voice, the actors were asked to address Sarah as if spoken by her friend when they are charged by a dangerous dog on a path in the woods. The actors were students or teachers of medicine, economics, logopedics, and engineering. At least three portrayals of each emotion by each actor were recorded, and these portrayals were used as stimuli in an emotion recognition listening test. Prior to their use as stimuli, the portrayals were first subject to a pre-selection listening test, whereby a small group of listeners selected the best portrayal from the three, and indicated whether they considered it to be a good representation of the target emotion category. Portrayals rated 'very bad' were excluded from the analysis, and portrayals by 4 actors who received multiple 'very bad' ratings were discarded from the study entirely. For the emotion recognition listening test, 73 listeners, principally students of medicine, psychology, and engineering, as well as university personnel, were asked to assign one of the 10 emotion categories to each sample they heard. 50% of the samples were identified correctly (chance being

10%). The ‘frightened’ portrayals were one of the most correctly identified categories, with samples correctly identified 64% of the time. From the acoustic analysis, ‘frightened’ portrayals showed increases in intensity, duration, and F0 mean when compared with a reference, neutral ‘naming’ sample.

Sobin & Alpert (1999) investigated acted portrayals of fear, anger, sadness and joy by female participants. They focused on non-professional female actors and female listeners because literature on non-verbal communication (Mehrabian (1972), Zuckerman et al.(1975)) suggested that female listeners are more accurate. Compared to other studies of acted emotion, the authors were able to recruit a large number of actors (31 female subjects aged between 18 and 35 years old). They were audio recorded, each reading out four sets of five stories that all included the sentence “it’s hard to believe this is real, I can’t believe things like this happen” in 4 different emotional contexts: fear, anger, sadness and joy. The sentences were then rated for emotional intensity by 12 female listeners, 3 per emotion, each of which listened to a range of produced sentences and rated them for their assigned emotion. 38 samples were then chosen to represent each emotion. These were sentences that were consistently rated as high on the target emotion and low on all other emotions.

Sobin & Alpert (1999) then conducted an acoustic analysis to find the features most closely correlated with the emotion scores. They found fear to be characterised by high pitch, high pitch variance and fast rate of utterance with few pauses. Fear and anger had similar acoustic profiles except that anger was associated with low pitch and high volume variance. Fear and joy/sadness shared no major characteristics.

Sobin & Alpert (1999) asked people producing the speech to self-rate after uttering each sample. In 97% of the stories, the actors reported feeling the target emotion while reading. However, they were unable to assess the degree to which they expressed the emotion effectively. When analysing listeners’ perceptions, Sobin & Alpert (1999) found that single parameters did not characterize the emotions, but rather listeners associated acoustic patterns consisting of a combination of parameters for characterisation. Of particular interest is the fact that listeners rating for fear had a high rate of false positives in sentences that were not produced in a

fearful context. Sobin & Alpert (1999) argue this may be because of an evolutionary advantage to over-detection of fear as opposed to under-detection.

Belin et al. (2008) ran a study of acted emotion that stands out in that it aimed to elicit non-verbal productions rather than emotional speech. 22 actors (11 male and 11 female), all French Canadian, were recorded producing a series of /ɑ/ vocalisations corresponding to happiness, sadness, fear, anger, pleasure, pain, surprise, disgust and neutral. These were then presented to 30 listeners who evaluated all productions on ten scales: one for each of the eight emotional categories, one for the valence of the actor's emotion (from negative to positive), and one for the perceived level of arousal of the actor.

Similarly to the other studies discussed, they found that in general recognition accuracies were high, with fear being correctly recognised 68% of the time (the highest recognition rate was for happiness at 81%, and the lowest for pain at 58%). The most common error involving fear was confusion with surprise. They also revealed significant effects of both the actors' and the participants' gender: the highest hit rates (75%) were obtained for female participants rating female vocalisations, and the lowest hit rates (60%) for male participants rating male vocalisations.

The productions of the 5 male and 5 female actors who received the highest scores overall were then analysed acoustically. They found that fear elicited a higher F0 than neutral productions (the mean F0 for fear vocalisations across all ten actors was 508 Hz, compared to 168Hz for neutral productions). This was the highest median F0 reported for any emotion. The maximum F0 for fear varied greatly, with the mean maximum being 642Hz, but the highest individual F0 being 1658Hz. On the other hand, there was no difference in intensity between fear and neutral productions.

Spackman et al. (2009) investigated performances of American trained and untrained actors, who had been instructed to read a standardised text in one of four emotional manners: fearful, angry, sad and happy. The 8 trained actors (4 male and 4 female) were all seniors (final year undergraduates) majoring in a fine arts course and had at

least 4 years professional theatre training, whereas the 8 untrained actors (4 male and 4 female) were introductory level psychology undergraduates with no professional theatre training. Both the trained and untrained actors were recorded reading the standardised text twice per emotion. The recordings were played to 215 listeners who were all undergraduate psychology students, roughly balanced for gender. The listeners were randomly assigned to groups, each group listening to one recording from each actor, divided in a way such that all four emotions were represented. After listening to each recording the listeners selected which emotion was being expressed from the 4 emotion category choices plus an "I don't know" option.

It was found that trained speakers' portrayals of anger and fear were more accurately identified than untrained speakers' portrayal of anger and fear. In contrast, untrained speakers' portrayals of happiness and sadness were identified more accurately than those of trained speakers.

Following the listening exercise, Spackman and colleagues conducted an acoustic analysis of the target sentences, which was normalised across speakers in order to ignore speaker variation, though it is not stated how this was implemented. They were interested in whether emotions have multiple or single acoustic profiles. They found that, generally, trained speakers spoke slower when portraying fear (and happiness) than untrained speakers, but vocal profiles for the emotions differed across speakers. Regardless, when the listener's rankings were considered, it was clear that despite this difference in trained and untrained profiles for fear, they were both perceived as fear. Therefore, variability amongst speakers did not affect accuracy of interpretation. Spackman et al. (2009) suggest that this may have been because their selection procedure meant that the trained and untrained were more similar to each other than in other studies that chose more experienced actors for trained conditions.

Spackman and colleagues criticise prior literature such as Banse & Scherer (1996) and Scherer (2003) for considering trained speakers to be better because they make so-called "cleaner" emotions, which assumes that there are basic (or discrete) emotions. However, this study shows that untrained speakers are just as capable of producing recognizable emotional portrayals, even though they are not as consistent.

Although acted emotions allow for greater control in terms of data collection, there is a risk involved with using non-naturalistic data, as the actors may be emphasising the cues that they are most aware of, thereby leading to a circularity of reasoning - in other words, participants may be finding it easier to identify emotions because the actors are using strategies aimed at maximizing identification, rather than replicating authentic portrayals which may be more ambiguous. Furthermore, the task of explicitly being asked to identify emotions can itself draw the perceivers' attention to cues that they might otherwise miss. In addition, as highlighted by Scherer (1986: 144):

> *It cannot be excluded that actors overemphasize relatively obvious cues and miss more subtle ones that might appear in natural expression of emotion.*

Therefore, it is not clear whether these results are fully representative of the emotion under investigation.

2.4.2 Experimentally-induced emotion and stress

Contrary to acted portrayals of emotions, these studies are designed to elicit genuine instances of the emotion from their participants, but unlike the studies discussed in the following section, they do so under laboratory conditions. Because it is difficult, and in many cases unethical, to induce actual distress, most induced studies focus on either cognitive or situational stress. Cognitive stress is usually induced by making the subject solve a puzzle or other difficult task, such as in Scherer et al.(2002) who recorded English, French and German speakers reading out sentences while performing logical reasoning tests, in two conditions, with and without distraction. They found that speech rate increases and F0 increases for high cognitive stress conditions compared to low stress conditions.

A study that compared the effect of cognitive stress with other negative emotions was Tolkmitt & Scherer (1986). Their participants were German university students who were selected to exhibit either 'low anxiety' or 'high anxiety' or 'anxiety-denying' personalities. The students were recorded while viewing a slide show

containing gruesome images of injured individuals as well as logical puzzles. For the gruesome slides, the participants were asked to say out loud what they thought the chance the injuries depicted would heal (expressed in a numerical scale within a carrier phrase). For the cognitive test slides they had to solve the puzzle and say it aloud, also within a carrier phrase. In other words, the task was designed explicitly to induce stress within the participants, either emotional stress (in the gruesome slides) or cognitive stress (in the puzzle slides). What Tolkmitt and Scherer found was that mean F0 was not affected by the manipulation, but that for high-anxiety and anxiety-denying subjects, the minimum F0 increased in stressful conditions. There were also significant changes in formant frequency values for female participants, with anxiety-denying females showing more precise articulation in cognitive stress conditions, but not during emotional stress.

Situational stress can be induced by asking subjects to perform a task they find socially stressful, such as deceiving an interviewer or falsifying their responses. Streeter et al. (1977) recruited male students to participate in an interview about their opinions. They were instructed to falsify responses to specific questions, but to answer truthfully to others. They found that fundamental frequency increases for lying (the stressful condition) as opposed to telling the truth. Similarly, Kirchhübel & Howard (2013) recorded male British English students in a mock theft paradigm in 'truth' and 'lying' conditions. The participants were instructed to deceive a 'security guard' (an experimenter) who would accuse them of stealing two objects. As part of the experiment, the students had stolen one item, producing speech under deception, but not the other, thus producing 'truth' speech. Unlike Streeter et al. (1977), they did not manage to find a correlation between deceptiveness and any acoustic feature examined (F0, intensity, and vowel formants F1, F2 and F3).

Another type of situational stress is Fuller et al.'s (1992) study, in which American female students were recorded producing vowel articulations (/a/ and /i/) two weeks before, one day before, and one day after their nursing exams. The students were classified for stress-coping style in groups of high anxiety, low anxiety and anxiety-denying students (similar to Tolkmitt & Scherer (1986)), according to their response to two anxiety questionnaires. The authors investigated some acoustic parameters

(including F0 and jitter) and some physiological parameters (including heart rate and sweating of the hands). Of interest here is that F0 was not found to be a reliable indicator of stress. Jitter, however, differed across both vowels in the 3 contexts, being greater the day before the exam, and did not differ across coping styles.

In a similar vein, Sigmund (2006) created the ExamStress database to study psychological stress in which of 31 male Czech students were recorded before, during and after their final oral examinations. In acoustic analyses, he found increases in F1, F2 and F0 mean and range in stress conditions when compared to neutral (post-exam) conditions.[4]

Several studies investigating speech stress are designed so as to examine combinations of cognitive, physical and situational stress. Hicks (1979) examined the speech of American male and female subjects who experienced electric shocks (lab-induced physical stress) and who had to deliver a speech in public (situational stress). He found that stressed speech (in either condition) led to changes in terms of decreased intensity, increased fundamental frequency, a decrease in speech rate, and longer speech bursts, but these changes were mainly significant when they occurred in the situational stress condition. He concluded that the type of stress determined the extent of the changes in stress speech versus 'normal' speech. He also observed that the variability of subjects meant that the changes were not uniform across subjects.

Jessen (1997) recorded 20 male German police officers in non-stress, physical and situational stress, and cognitive stress conditions. 10 of the subjects had little training for shooting firearms in real-life situations and were considered 'untrained' in terms of extreme stress management. The other 10 subjects formed part of a 'trained' group who had been trained to deal with hostage situations and terrorist attacks. All the police officers were subject to a maths test while listening to audio from a hostage situation (cognitive stress), followed by a physical task of shooting specified targets with live ammunition while answering a maths problem (situational and

[4] Fuller et al. (1992) and Sigmund (2006) are classified as induced stress studies here, even though the exam situations were genuine, as they share the core property of laboratory studies in that subjects were pre-selected and stress conditions were planned for.

physical stress). Jessen found that F0 mean and standard deviation increased when produced in stress conditions, and the extent of the increases was greater in the physical/situational stress condition than in the cognitive stress condition. The untrained subjects exhibited F0 changes in the cognitive stress condition whereas the trained did not, but both groups showed such changes for the physical/situational activity and these changes were greater in the trained subjects. Jessen does suggest that the reason trained subjects exhibited changes in the physical/situational condition may have been due to the exertion involved in the task and not necessarily an indicator of stress. The untrained group, however, did not engage as much in the physical task and therefore their results are argued to reflect genuine situational stress. Therefore, these results resemble those found by Hicks (1979) in that for the subjects that display both, situational stress has a greater effect on speech than does cognitive stress.

Meinerz (2008) recruited 30 German subjects (15 male and 15 female) to investigate situational stress in the form of a simulated job interview, as well as cognitive stress by way of solving a maths problem. He analysed F0 mean, median and standard deviation, articulation rate, syllable rate and formants F1, F2 and F3. He found no significant changes in F0 across the stress conditions. F1 and F3 increased under cognitive stress conditions, but articulation and syllable rate increased under situational stress conditions (which is a somewhat different from what Hicks (1979) and Jessen (1997) found above).

Sometimes, induced emotion studies further manipulate the expression of the emotion. Ekman et al. (1976) investigated the use of F0 and hand gestures in deceptive speech (though only F0 will be reported here). They recorded 16 American student nurses in two types of standardised interviews in which they were to describe their feelings and emotional state after having viewed either pleasant or gruesome images. In the ‘honest’ interview, the nurses viewed the pleasant images and were asked to describe their feelings honestly. In the ‘deception’ interview, they watched gruesome slides of burns victims and amputees, and they were requested to disguise their initial reaction and instead maintain and describe their persona as positive. It

was found that the F0 increased in the nurses' speech produced in the 'deception' interviews.

Ekman et al. (1976) then ran a perceptual experiment in which the interview recordings were played to groups of observers. The observers were played both honest and deception interview samples from different speakers, and were asked to rate the samples on 14 bipolar scales. The observers were divided into four groups. Some saw a video showing the faces of the speakers as they talked, others saw a video showing the body with the face cut out, and two groups heard audio-only recordings. Among those, the first group heard the audio of the interview without any filtering, and the second heard the interview with all frequencies above 400 Hz removed (making the speech unintelligible). Of interest to us is the fact that among those who heard the audio only (unaltered or filtered), there was no effect on observer rankings of whether the speech was honest or deceptive.[5]

As can be seen from the above examples, one of the problems with induced studies is that it is not clear how genuine the emotion induced actually is. While these experiments do allow the experimenters a level of control over what the participants experience, this control is limited, as there is no way to guarantee that the participants really are emotionally invested in the experimental results. Furthermore, it has been argued that experimentally-induced emotion often results in a weak portrayal of the emotion (Scherer 2003).

There was one attempt to compare induced emotion directly to acted emotion. Wilting et al. (2006) who, following a methodology first introduced by Velten (1968) for mood manipulation, recorded Dutch academics reading positive, negative and neutral sentences while instructed to display positive and negative emotions. Wilting and colleagues believed that displaying emotions may lead to feeling them, and therefore considered positive sentences read in a positive way to be 'real' emotion, and positive sentences read in a negative way as 'acted' (and vice versa for negative sentences). Their perceptual study then compared real and acted responses

[5] The observers who saw video ranked the producers differently depending on whether they saw faces or hand gestures. However, these findings are not directly relevant to the topic of this thesis.

to positive and negative emotion among Dutch listeners. Stimuli were presented as audio-visual and audio only. Acted emotions were judged as stronger than "real" emotions, which led the authors to cast doubt on the use of actors as a way to study real emotions.

Barkhuysen et al. (2007) used the same recorded material as Wilting et al. (2006) but conducted the perceptual experiment using Czech listeners rather than native Dutch listeners. Overall, they replicated the original results, though differences between acted and non-acted emotions were found only in the audio-only condition.

2.4.3 Genuine emotion and stress

Genuine emotional data form a minority of the studies in emotional speech research since they are harder to collect and control. It is favoured in emotion studies due to its high ecological validity. However, instances of naturally-occurring emotion are often brief, and generalisations about them tend to be based on few observations. The quality of recording is also typically worse than those in more controlled conditions, reducing analysis options. Furthermore, it is often hard to determine the precise nature of the emotion being analysed. Researchers who have used genuine emotional data have typically analysed the speech of aviation personnel from flight recorders after emergency situations.

Williams & Stevens (1969) investigated speech produced by American pilots and control tower operators before and during flight difficulties, as well as the radio announcer who was broadcasting during the Hindenburg disaster. They analysed F0 mean, range and contour, and found that both F0 mean and range increased during and immediately after serious flight difficulties when compared to speech produced before the flight difficulties. They also found that the F0 contour often became irregular and discontinuous as the flight difficulties progressed. Other studies that have examined the speech of pilots during emergency situations include Kuroda et al. (1976), Brenner et al. (1983), and Ruiz et al. (1996), whose results indicated that F0 increased during the emergency situations.

Protopapas & Lieberman (1997) put aviation data to a different use as they tested the correlation between F0-based parameters in the speech of a male helicopter pilot and the perception of stress by listeners from a university community. They did so by combining synthetic vowel stimuli with the pilot's naturally-occurring F0 contours from stressed and non-stressed conditions. They found that excerpts with increased mean and maximum F0 were more likely to be rated as stressed, whereas F0 range did not correlate with the perception of stress. These studies are particularly relevant to the forensic context as they also involve individuals in mortal danger and are therefore comparable with the data analysed for the current investigation.

Studies of speech under stress also make use of induced and naturalistic emotional speech paradigms. The studies of speech during aviation disasters mentioned in the previous paragraphs are typically included in reviews of speech under stress literature as examples of psychological stress. Other studies using naturally-occurring speech under psychological stress speech include Streeter et al.'s (1983) analysis of recordings during the 1977 New York city blackout (which led to widespread disorder, looting and arson) between a system operator and supervisor of the power provider of the city. They found that the F0 produced by the supervisor increased with increased situational stress, but that there was an inverse correlation between the operator's F0 and his stress level. In a perceptual experiment based on excerpts from the two recordings, undergraduate student listeners perceived increases in F0 variability, mean F0 and amplitude as indicators of stress. A second study using naturalistic non-aviation data was conducted by Devillers et al. (2004) who investigated psychological stress in speech exchanges between French clients and agents of a web-based stock exchange customer service centre. They did not compare the absence or presence of stress, but compared two related emotions - fear and anger - to each other. They found that F0 variation and pause features can be reliable indicators of the two negative emotions, but are not very reliable in distinguishing between the two. They conclude that manifestations of fear and anger are variable and dependent on variables such as speaker role and gender.

Overall, the picture that emerges from studies of genuine stress data is inconclusive, with the majority of studies finding that mean F0 increases with stress but some

studies finding other F0-related measures affected instead. Furthermore, the role of the recorded individual in the conversation and the level of training s/he received in dealing with stressful situations has an effect on the manifestation of stress. Perceptual data, on the other hand, seems to indicate that F0 mean is always positively correlated with perception of stress, regardless of who produced the data.

It is worth noting that all genuine data fall under the general category of situational stress; unlike in stress-induced data studies, cognitive stress is not taken to be of interest here. This means that the majority of induced stress studies use a different baseline of comparison than genuine stress studies, making their results harder to compare directly.

2.5 Analysis of screamed productions

Relevant to the forensic context is the study by Begault (2008), who recorded and analysed American females, presumably non-professional actors, producing screams. The motivation behind his study was questions from attorneys concerning the audibility of female screaming.

Begault's study was based on an actual incident in which a female victim was alleged to have screamed during an attack; he attempted to replicate the scenario in order to assess how factors such as distance and obstructions interacted with audibility.

Begault recorded 10 female subjects, aged between their mid-20s and mid-40s, producing three screams each. All the subjects were instructed to scream "as loudly as possible, as if you had just been surprised by something very scary". The recorded screams ranged in intensity from 102-123 dB (mean 114 dB). Only the loudest scream from each subject was included in the analysis. Begault found a correlation between age and intensity, with the loudest screams produced by subjects under 30.
One recording (the second loudest scream overall, lasting 2.25 seconds) was selected for use in the second phase of the study (no clear criteria as to how it was chosen is provided). This recording was played on a loudspeaker at a specific indoor location (replicating the account of the actual incident). It was repeated four times. The measurements all occurred at the same time of the day, same day of the week, and

under similar weather conditions as those on the date of the incident. Instruments were used to see if the recording was perceptible at various locations around the source location – three outside the source location, and one indoors. The sounds at those locations, including background noises, were recorded. Begault's main finding was that the screams could indeed be detected in all measurement locations.

The findings in Begault's study cannot be directly compared to the current research as the only acoustic parameter involved was intensity. However, it is still of importance, as it is the only other study of screams from a forensic perspective, and it demonstrates the importance of scream research to forensic practice.

2.6 Summary of empirical results

As can be seen from the discussion above, the current literature gives rise to a complex, sometimes contradictory, picture of both the acoustic profile and perceptual results associated with fear and stress. This section will present this picture directly by summarising the results of all the studies described by measure.

2.6.1 Production

Table 2-3 below contains a summary of the data from all the studies for which acoustic analyses of fear/stress speech were conducted, organised by acoustic parameter. Three parameters were identified in more than one or two studies. Almost all studies found an increased F0 mean associated with both stress and fear speech, though some studies found an increase in a different F0-related measure (such as F0 maxima). Intensity was found to be increased for fear, but Hicks (1979) found that it decreased for stress; because this is a single data point, however, it cannot be determined whether this result is an outlier or whether intensity can be used to distinguish the two emotions. Finally, speech tempo was implicated in seven studies, but is highly inconsistent, with four studies finding that fear and stress speech display a decreased rate, and three finding an increase.

Table 2-3: Summary of emotional speech literature according to acoustic variable.

Variable	*Study*	*Type*	*Emotion*	*Finding*
F0	William & Stevens (1969)	Genuine	Stress	Increased F0 mean Increased F0 range Irregular F0 contour
	William & Stevens (1972)	Acted	Fear	F0 peaks had "unusual shape" Increased F0 maximum
	William & Stevens (1972)	Genuine	Fear	F0 peaks had "unusual shape" Increased F0 range
	Ekman et al. (1976)	Induced	Stress	Increased F0 mean
	Kuroda et al. (1976)	Genuine	Stress	Increased F0 mean
	Streeter et al. (1977)	Induced	Stress	Increased F0 mean
	Hicks (1979)	Induced	Stress	Increased F0 mean
	Brenner et al. (1983)	Genuine	Stress	Increased F0 mean
	Streeter (1983)	Genuine	Stress	Increased F0 mean (supervisor) Decreased F0 mean (operator)
	Tolkmitt & Scherer (1986)	Induced	Stress	Increased F0 minimum
	Scherer et. al (1991)	Acted	Fear	Increased F0 mean and variance
	Banse & Scherer (1996)	Acted	Fear (panic)	Increased F0 mean and energy
	Ruiz et al. (1996)	Genuine	Stress	Increased F0 mean
	Jessen (1997)	Induced	Stress	Increased F0 mean Increased F0 standard deviation
	Leinonen et al. (1997)	Acted	Fear	Increased F0 mean
	Sobin & Alpert (1999)	Acted	Fear	Increased F0 mean and variance
	Devillers et al. (2004)	Genuine	Fear & Anger	Increased F0 variation
	Sigmund (2006)	Induced	Stress	Increased F0 mean
	Belin et al. (2008)	Acted	Fear	Increased F0 median Increased F0 variation
Formants	Sigmund (2006)	Induced	Stress	Increased F1 and F2 mean
	Meinerz (2008)	Induced	Stress	Increased F1 and F3 mean
Jitter	Fuller et al. (1992)	Induced	Stress	Increased jitter
Speech tempo	William & Stevens (1972)	Acted	Fear	Decreased speech rate
	Hicks (1979)	Induced	Stress	Decreased speech rate
	Banse & Scherer (1996)	Acted	Fear (panic)	Increased speech rate
	Leinonen et al. (1997)	Acted	Fear	Decreased speech rate
	Sobin & Alpert (1999)	Acted	Fear	Increased speech rate
	Meinerz (2008)	Induced	Stress	Increased speech rate
	Spackman et al. (2009)	Acted	Fear	Decreased speech rate (for trained actors)
Intensity	Hicks (1979)	Induced	Stress	Decreased mean intensity
	Scherer et. al (1991)	Acted	Fear	Increased mean intensity Increased intensity variance
	Banse & Scherer (1996)	Acted	Fear (panic)	Increased mean intensity
	Leinonen et al. (1997)	Acted	Fear	Increased mean intensity
	Belin et al. (2008)	Acted	Fear	No change

2.6.2 Perception

While most studies described above included a perception component, the variation in methodologies and baselines used mean that there is no single metric that can be compared. Generally, the emerging picture from the literature reviewed is that people tend to recognise fear and stress at rates higher than chance, but that this is influenced by a variety of situational variables, as well as by who produced the speech and, more controversially, by whether the listener was male or female. This seems to be in accordance with findings in the more general emotional literature, such as those reported by Scherer (1989; 2003) and Johnstone and Scherer (2000) discussed earlier in the chapter. It is also worthy of note that although perceptual tests can augment findings from production studies, they can be criticised for testing emotion discrimination, rather than recognition (Scherer, 2003:234). This means that listeners will select from alternatives, rather than choose an emotional category in its own right. This is problematic when interpreting emotion accuracy rates.

2.7 Chapter summary

This chapter summarised the state of the field of emotional speech research, with particular reference to studies investigating production and perception of fear and speech under stress. A brief history of emotional speech research was provided, followed by a physiological description of voice and emotion. Advantages and disadvantages of different emotional speech paradigms were considered, and the relevant studies discussed in detail. A summary of principal findings from relevant empirical studies were presented, showing that a complex, somewhat inconsistent picture emerges.

3. Methodology of Acoustic Study

This chapter contains a description of the methodology of the first of two principal studies conducted for this thesis: the acoustic investigation of acted and authentic distress. (The methodology of the second study, that of the perceptual investigation of distress, can be found in chapter 6). The first section signposts the research design of the study and defines the type of speech and speaker under investigation. The collection and background of two corpora are explained in the next section. An overview of the parameters chosen for analysis and a description of analytic techniques and equipment are provided in the third section. The final section summarises the chapter.

3.1 Research Design

The first research question presented in §1.2 - to what extent can specific acoustic measures be used to identify distress speech? - is addressed by conducting an acoustic investigation of recorded distress. Two data sets are used for analysis. The first comprises authentic forensic recordings which contain productions by real victims of violent attacks. Analysing them involves working inductively from the speech material using auditory-acoustic analyses. The second is a data set of recordings of speech produced by professional actors. Vocal parameters of interest in the acted data set are compared and contrasted with those of victims in the authentic dataset.

As illustrated in Chapter 2, the majority of emotional speech studies have examined vocal cues to either laboratory-induced or acted emotion. Few attempt to examine extreme emotion and fewer use authentic data (though there are some notable exceptions in the form of studies reporting on stress in the speech of aviation personnel, §). Most existing studies, although insightful, have the following limitations:

1. they employ a simplistic categorisation of every-day emotion, e.g. ‘happy’, ‘sad’, ‘angry’;

2. they have limited ecological validity (i.e. they are typically not applicable to real-life situations);

3. they may reflect stereotyped or stylised states, i.e. artificial modes of behaviour, that actors are trained to adopt.

This PhD investigation represents a small first step towards redressing these limitations by analysing both authentic and acted distress responses. Distress represents an extreme emotion, easily conceptualised but hard to define. It is one that is not found in every-day interactions, but is frequently prevalent in forensic situations. Speech and vocalisations from victims in real-life emergency situations are compared with corresponding reference (non-distress) material from the same speakers. In addition, by contrasting these with productions of actors attempting to enact the same situations, the study investigates the relationship between authentic distress and potential stylised emotional behaviours.[6] Furthermore, the investigation benefits from data collected in a more controlled environment that may allow for clearer observations between distress and non-distress speech. By combining carefully controlled laboratory distress material with real-life 'messy' forensic data - i.e. recordings that are frequently of the brief duration and variable quality that is typical of data used by forensic practitioners - the present study aims to contribute towards results that might ultimately be used to substantiate forensic expert opinions in this area.

3.2 Data

Two corpora have been collected and are described below. The first corpus comprises eight authentic forensic cases and the second contains acted data from six comparable forensic scenarios.

3.2.1 Authentic data

One consequence of the widespread availability and use of mobile telephones is that many violent attacks are now audio recorded, as victims and witnesses to such

[6] Reference material from victims and actors is not compared since the focus of the investigation concerns distress speech within and across speakers, namely actors and victims. Furthermore, reference material from victims, although occasionally available, is rare. Reference material from victims is analysed where possible, but it is harder to obtain than reference material from actors.

crimes (and at times the perpetrators themselves)[7] often telephone the emergency services when an attack is imminent or in progress. Examples of recordings from past criminal cases are held in an archive at JP French Associates (JPFA), a laboratory specialising in the forensic analysis of speech, language and audio. An agreement was made that recordings from case material relating to violent crimes be made available to me for the purposes of the present research.

Cases in which distress vocalisations were known to be present in the recording (e.g. through police reports detailing the violent attack, injury, and in some cases the death of the victim) were selected for analysis. Initially, eight cases were selected but, owing to the brevity of some data, two cases were omitted from the acoustic analyses (although these were later used in the small-scale perceptual experiment described in Chapter 4). The victims had been subjected to a physical attack, typically involving a knife or gun, and were in need of emergency assistance at the time of the recording.

All victims originated from either England or Wales and spoke English as their native language. Despite the difficulties in controlling for a balanced sample of victims of violent attacks with regard to social and demographic information,[8] the data were, on the whole, balanced in terms of including the voices of victims of different sexes and ages, with four of the eight original recordings containing male victims aged between 17 and 47 years old, and the other four recordings containing voices from female victims aged between 15 and 31 years old.

[7]Examples of the attacker (posing as a witness) calling the emergency services on behalf of the victim are found in some forensic casework. Moreover, the recent 'happy slapping' phenomenon, whereby someone assaults an unwitting victim while s/he, or an accomplice, records the assault (typically on a camera phone), has been widely reported in the media, e.g. in the BBC (Akwagyiram, 2005) and in The Guardian (Honigsbaum, 2005).

[8] The archive at JPFA contains a range of recordings from past criminal investigations, but not all are suitable for this research since: (i) not all recordings concern violent attacks, (ii) where violent attacks have taken place the recording does not always contain material uttered by the victim, (iii) if the victim's speech is present, the material is often brief and/or overlaps with speech from other speakers. The authentic material used in the present investigation is analysed based on what is available for analysis, and not according to any preconceived selection criteria.

A brief outline of the circumstances of each case, as well as the material available for analysis, is set out below. Where possible, reference data, such as recordings of some victims in non-distress circumstances, were obtained. For two victims, both males, reference data was available, as speech material had been recorded prior to attack. Due to the sensitive nature of the material, all data are anonymised with references to the victims' and perpetrators' identities removed, following standard ethical procedures. Victims are not referred to by name. Rather, a letter is assigned to them, e.g. 'Victim A', and an additional speaker code is given in the format of initial:number(s):initial. The first set of initials refers to the case letter. (N.B. it is not the initial of the victim). The numbers refer to the victim's age. The final initial denotes the victim's sex as either M (male) or F (female). For example, A34M refers to the victim from case A. The male victim was 34 years old at the time of the recording. For readability, when referring to a victim, the letter should suffice (e.g., Victim A) though the speaker code, e.g. A34M, may be used in addition from time to time to remind the reader of the victim's age and sex. It should be noted that none of the below cases is 'live', i.e. still in legal dispute. All have already been resolved through the courts.

3.2.1.1 Case A

This case consists of an audio-recording of Victim A (speaker code A34M) in conversation with two other men, also aged between 30 and 40 years, in a car. Unlike all the other cases, this recording is not from a call to the emergency services. The car contained a recording device which was able to transmit events to a remote control centre, e.g. as seen in taxis or emergency service vehicles. All three participants were aware of this device. Towards the end of the recording one of the men drew a concealed gun and fired a series of shots. The shots were directed at the two other men as they attempted to get out of the car, as well as at another man in the immediate vicinity. There was no indication at the beginning of the recording that the man had a gun or would be violent. In fact, the victims thought the gunman was feeling ill and were talking to him about this.

The audio recording contains 14 minutes and 10 seconds of material, most of which was comprised of conversation between the participants prior to the violent incident, i.e. in a non-distress environment. There are also some minutes at the end of the

recording in which passers-by can be heard trying to help the victims after the perpetrator had fled. There are 22 seconds of edited, non-overlapping, pre-attack (i.e. non-distress) speech from Victim A that is suitable for acoustic analysis. Towards the end of the recording (from 11m 37s), 11 seconds of material are available starting from the moment whereby all participants become aware that a gun is present and about to be fired, and ending with the fifth and final shot, after which the attacker fled. Throughout these 11 seconds, the speech of Victim A is especially clear, presumably due to his location nearest the recording device, providing genuine distress speech which can be compared with his non-distress speech from earlier in the recording.

3.2.1.2 Case B

This case consists of a 999 call made from a mobile telephone outdoors by a man in his early forties, Victim B (B42M). The victim had been stabbed in the stomach following a drugs dispute and towards the end of the recording a vehicle can be heard approaching, and then knocking him over. The phone falls to the ground but speech from the victim and further vehicle noise is audible. The original call lasts 4 minutes 42 seconds. However, only the first 96 seconds contains useful speech material from the victim. After the victim has been knocked over, the line remains open but the recording contains mainly speech from the operator. There are 22 seconds of edited, non-overlapping speech material of Victim B leading up to the car attack, and then a further 11 seconds where the victim can be heard overlapping with the car noise and the operator.

Reference data are available for Victim B from a previous criminal incident in the form of a police custody interview[9] lasting 6 minutes and 23 seconds. The interview provides mainly 'no comment' style responses, though 20 seconds of edited, non-overlapping speech is available.

[9] Following the Police and Criminal Evidence Act 1984 (PACE) Code E, all suspects in England and Wales are recorded during their police custody interviews.

3.2.1.3 Case C

The material from this case is a 35-second recording of a call from a landline telephone by a man, Victim C (C47M), to the emergency services after he had been stabbed in the neck by an acquaintance when at home. The victim had barricaded himself in a room of the house following the attack, and then had called the emergency services. The victim's conversation with the operator ends abruptly; it remains unknown whether the victim experienced a further attack towards the end of call. There are 14 seconds of edited, non-overlapping speech available for analysis. No reference data are available for this victim.

3.2.1.4 Case D

This case involves a call to the emergency services, lasting 34 minutes 33 seconds, made from a mobile phone. Victim D (D15F) was a 15-year-old girl who witnessed the shooting of a close friend in her house by a family member. She called the emergency services while trying to run and hide from the attacker who was still present in the house. The first 2 minutes and 20 seconds of the recording contains evidence of an initial struggle between the attacker and Victim D, as well as repeated attempts by the attacker to gain access to the victim's location, before a final struggle can be heard resulting in further shots being fired. The rest of the call contains the operator's speech as she tries to re-establish contact with the victim before police officers arrive. 36 seconds of non-overlapping speech is available for analysis. No reference data are available for this victim.

3.2.1.5 Case E

Victim E (E17F) was a 17-year-old girl calling the emergency services from her landline phone late at night after having been accidentally shot in the chest by a friend in her home. The call lasts 34 minutes and 1 second. Only the first 72 seconds contain conversation between the victim and operators. Due to overlap between speakers, only 8 seconds of speech from the victim are useable for analysis. The rest of the call records the ambulance service call taker giving first aid advice to a friend of the victim at the scene until the ambulance arrives. There are no reference data for this victim.

3.2.1.6 Case F

This case involves a recording from a landline to the emergency services by a woman in her late twenties after she had been shot in the head. Victim F (F27F) was at home when a thief entered the house and shot the victim's partner and then the victim. While the house is being burgled, Victim F calls the emergency services. The attacker hears her and returns to shoot her. The call lasts 8 minutes and 55 seconds, but only the first 33 seconds contain vocalisations by the victim. The operator mistakes the victim's cries for speech from a child. The line is left open but the remainder of the recording contains only background noises of the burglar in the property. 9 seconds of the victim's edited, non-overlapping speech are available for analysis. There are no reference data for this victim.

3.2.1.7 Case G

Case G is a call made from a mobile phone outdoors by a woman who is reporting an attack on Victim G. G31F was heard in the background moaning and wailing after having sustained severe head injuries while the caller and her boyfriend, the perpetrators of the attack, direct the emergency services to their location. The call lasts 2 minutes and 36 seconds, of which 15 seconds contains vocalisations from the victim. No reference material is available.

The recording from this case does not form part of the acoustic analysis and is therefore not compared with the acted distress responses. This is partly due to the fact that the victim is distant from the microphone, and partly because the victim's responses did not lend themselves to being part of a script that could be later performed by actors (see §3.2.2.3). However, there are some occasional vocalisations that do not overlap with the speech of the caller and operator, which were later used in the small-scale perceptual experiment described in Chapter 4.

3.2.1.8 Case H

This case involves a call made by the victim, H17M, from a mobile phone outdoors requesting an ambulance after he has been attacked with an unidentified blunt instrument by a group of acquaintances. The call lasts 1 minute but only 3 seconds of the victim's speech are suitable for analysis since one of the attackers quickly takes over the phone before the victim can tell the emergency services where he is. The

attacker informs the operator that he does not know where they are but will call back once he knows their location. No other call is made to the emergency services and the victim is further attacked. There is no reference material available for this victim.

As with Case G, this recording does not form part of the acoustic analysis due to the brevity of the victim's speech. However, it was used in the small-scale perceptual experiment described in Chapter 4.

3.2.1.9 Case material summarised

Table 3-1 on the opposite page provides a summary of the case circumstances and recordings used.

Table 3-1: Summarised case information relating to authentic forensic material

VICTIM	SPEAKER CODE	CIRCUMSTANCES	TYPE and DURATION of RECORDING	DURATION of SAMPLE
A	A34M	The victim is shot suddenly and unexpectedly by a car passenger.	1x recording in 2 parts using recording device (total 14m 10s) i) Non-distress conversation between passengers ii) Distress speech from victim	reference = 00m 22s distress = 00m 11s
B	B42M	The victim has been stabbed (outdoors) in the stomach following a failed drug sale and is later run over by a car.	2 x recordings: 1 x mobile phone 999 call reporting stabbing (04m 42s) 1 x police interview (reference speech) (06m 23s)	reference = 00m 20s distress = 00m 33s
C	C47M	The victim has been stabbed in the neck after an altercation with an acquaintance at home	1 x landline phone 999 call (00m 35s)	00m 14s
D	D15F	The victim has witnessed the shooting of a friend and hides in the house while the attacker chases her then shoots her.	1 x mobile phone 999 call (34m 33s)	00m 36s
E	E17F	The victim is shot in the chest accidentally by a friend when returning home.	1 x landline phone 999 call (34m 01s)	00m 08s
F	F27F	The victim witnesses the shooting of her husband before being shot herself during a burglary.	1 x landline phone 999 call (08m 55s)	00m 09s
G	G31F	The victim has been beaten up and sustains severe head injuries. She can be heard in the background.	1 x mobile phone 999 call (02m 36s)	00m 15s
H	H17M	The victim is in the process of being beaten up.	1 x mobile phone 999 call (01m 00s)	00m 03s

3.2.2 Acted data

To collect data from actors portraying victims in distress, a drama workshop was held at a professional recording studio in London in July 2010. The workshop was organised and run by myself in collaboration with Morwenna Rowe, voice coach and founder of Speak Easily, a company providing voice, speech and accent training to professionals, and Dan Barnard of Living Pictures Ltd, a training company for actors and directors. The latter two parties were interested in the portrayals of extreme emotion from a dramatic training perspective and were on hand throughout the day to assist me in my leading of the workshop and to ensure that actors did not risk damaging their voices and mental well-being when working on distressing material.

3.2.2.1 Stimuli

Actors were recorded performing scripts based on events in cases A-F from the corpus of genuine forensic cases (described in the previous section) during the workshop. They were presented with a brief written case background followed by a script based on an amended transcript of the original case material (Appendix A). Personal information and personal relationships were changed so that the real victims' identities could not be worked out at a later date using, for example, internet search engines. Where potentially identifying information was changed, new words were created while retaining, where possible, the same number of syllables, a similar stress pattern, and a similar phonological content, especially in terms of stressed vowels. For example, if (hypothetically), the real victim revealed her name and address in the authentic recording to be Traci Lane from Nine Foss Street, the new script could be Gracey Staines from Five Moss Street. As some cases received strong media attention and were published and reported on nationally, I changed some of the personal relationships within the case background while still maintaining a similar relationship dynamic. For example, if the original forensic case concerned an attack on a teenage boy by his violent step-mother, the case background might state that the actor is playing a male adolescent who was attacked by his aunt. Furthermore, actors were only informed of the approximate age of the victim, e.g. teenager, twenties to thirties, or forties to fifties.

3.2.2.2 Actor participants

Twelve actors (six male, six female) were recorded in total. Two actors worked on each script and kept with the same script for the entirety of the workshop. Scripts were assigned to actors randomly, albeit taking into account the sex (and where possible the age) of the victim. The actors were not exposed to the original case material prior to the recording sessions. The actors were aged between 23 years and 48 years (male mean age = 29 years, median = 27 years; female mean age = 29 years, median = 28 years), and all were self-reported speakers of Standard Southern British English, with the exception of one female actor (Actor 10) who spoke Irish English. They had all completed a National Council Drama Training-accredited programme in the UK and were recruited through drama mailing lists.

The actors are not referred to by name but by number, and were assigned a speaker code in the same way as the victim speaker codes. However, for the acted data, the first initials do refer to the actors' initials, unlike the victim speaker codes. Thus 'JS23M' refers to an actor with the initials JS, who is 23 years old and male.

3.2.2.3 Elicitation and recording of acted distress speech

The actors underwent a series of warm-ups and rehearsal exercises chosen by Morwenna Rowe and Dan Barnard. The exercises were based on a psycho-physical approach to acting, pioneered by Konstantin Stanislavsky (see, e.g. Stanislavsky (1968), Benedetti (2000), and Merlin (2001)). In a psycho-physical approach, there is an emphasis on playing action since it is believed that through action, actors can contact their emotions (Merlin 2014: 135). Otherwise, there is a danger of generalising the emotion (Merlin 2014: 135), which may lead to an inauthentic performance (Morwenna Rowe, p.c.). For example, in a scenario where an actor is called upon to portray a fearful hostage, s/he may portray their fear by attempting to befriend the hostage-taker in order to survive the ordeal (Merlin 2010: 65).

Since actors are sometimes requested to produce emotion at an extreme level from 'cold' (i.e. with little preparation or rehearsal time), especially in TV and film, it was established that the workshop sessions should allow both the directors and actors to explore playing extreme emotion from both 'cold' and 'hot' (i.e. with preparation

and rehearsal time that included vocal, physical and emotional warm-ups and exercises). The schedule was developed in conjunction with Rowe and Barnard to serve the needs of this investigation while best addressing the realities the actors experience. A schedule of the workshop can be found in Appendix B.

Each actor was recorded in three conditions: 'reference' (baseline speech material), 'unrehearsed' (known as 'cold' to the participants in the workshop), and 'rehearsed' ('hot'). The reference and unrehearsed speech material were collected as part of the same recording session, in which the actors were first asked to read the first half of the standardised reading passage 'The Rainbow Passage' (Fairbanks, 1960: 124-139) as control speech material (reference material). They then performed their script with minimal vocal, physical and emotional rehearsal (unrehearsed material). For the rehearsed condition, the script was performed and recorded later in the day after having undergone a series of vocal, physical and emotional warm-ups and exercises (rehearsed material). Recording the actors at three stages of rehearsal - reference, unrehearsed and rehearsed - was designed to allow for the observation of potential gradual changes between reference and distress material in the acoustic signal, and to elicit intermediate, though not necessarily equidistant, points on a continuum.

The scripts all contained the victim interacting with another interlocutor, typically the emergency services operator. An additional trained director, also present for the workshop, played the part of the interlocutor for all recordings. Cases which had originally featured three or more participants in the recording were edited so that only the victim and the operator (or, for Case A, the victim and attacker) were left in the script. Cases G and H were excluded as possible stimuli, as adapting their scripts would have been problematic given that the main participant in the call, alongside the operator(s), was the attacker and not the victim. The actors were recorded individually in a small room with corkboard flooring and minimal furniture away from the main rehearsal space.

Recordings were made using a head-mounted DPA 4066 microphone and a Marantz PMD 670 solid state digital recorder (44.1 kHz, 16 bit). The head-band microphone was positioned a few centimetres from each actor's ear, avoiding direct contact with

the skin. Although it is usually recommended that the microphone be worn nearer the mouth, screamed sections produced during equipment test performances resulted in clipping of the waveform due to signal overload. Before performing the script, actors were recorded producing a sustained vowel in order to calibrate intensity readings using a sound level meter. Performances were also filmed using a Panasonic S.D.R-H90 video camcorder, though video analysis does not feature in the current investigation. A full list of equipment used for recording the actors is provided in Appendix C.

Due to a technical error in Actor 7's first recording session, her 'unrehearsed' performance was not recorded correctly and therefore is not available for analysis. However, recordings of her reference speech material and 'rehearsed' distress performance were recorded correctly and have been included in the investigation.

3.2.2.4 Acted material summarised

Table 3-2 below contains a summary of the acted material and shows which scripts, based on the real-life scenarios of the victims in §3.2.1, were performed by which actors.

Table 3-2: Summarised information of acted material

CASE	VICTIM CODE and INFO	ACTOR	ACTOR CODE	DURATION of SAMPLE
A	A34M - English victim, shot in a dispute	1	JS23M	reference = 00m 50s unrehearsed = 00m 05s rehearsed = 00m 11s
		2	PB23M	reference = 00m 50s unrehearsed = 00m 04s rehearsed = 00m 05s
B	B42M - English victim, stabbed and run over by a car following a dispute	3	RG27M	reference = 00m 52s unrehearsed = 00m 36s rehearsed = 00m 28s
		4	PW28M	reference = 00m 51s unrehearsed = 00m 31s rehearsed = 00m 43s
C	C47M - English victim, stabbed in the neck following a dispute	5	MS27M	reference = 00m 40s unrehearsed = 00m 17s rehearsed = 00m 17s
		6	TB48M	reference = 00m 46s unrehearsed = 00m 17s rehearsed = 00m 20s
D	D15F - English victim, shot following a dispute	7	ZR28F	reference = 00m 40s unrehearsed = n/a rehearsed = 00m 49s
		8	ZC30F	reference = 00m 54s unrehearsed = 01m 01s rehearsed = 00m 55s
E	E17F - Welsh victim, accidentally shot by a friend	9	SS25F	reference = 00m 57s unrehearsed = 00m 12s rehearsed = 00m 51s
		10	DM27F	reference = 00m 40s unrehearsed = 00m 13s rehearsed = 00m 14s
F	F27F - English victim shot twice by an intruder	11	SS28F	reference = 00m 43s unrehearsed = 00m 03s rehearsed = 00m 08s
		12	TM38F	reference = 00m 45s unrehearsed = 00m 06s rehearsed = 00m 06s

3.3 Analysis techniques

3.3.1 Preliminaries

All authentic forensic recordings, with one exception, were received from JPFA in a .wav file format. These were digitised versions of the original unedited recordings. Due to the poor quality of the original Case D recording, however, an enhanced version to which a digital band-pass filter had been applied was also provided. This enhancement was originally performed in order to facilitate impressionistic listening and reduce listener fatigue. The main effect of the filter was to suppress extraneous noise, at frequencies below 100 Hz and above 3.6 kHz, which was presumably incorporated during the re-recording process of the original call.[10]

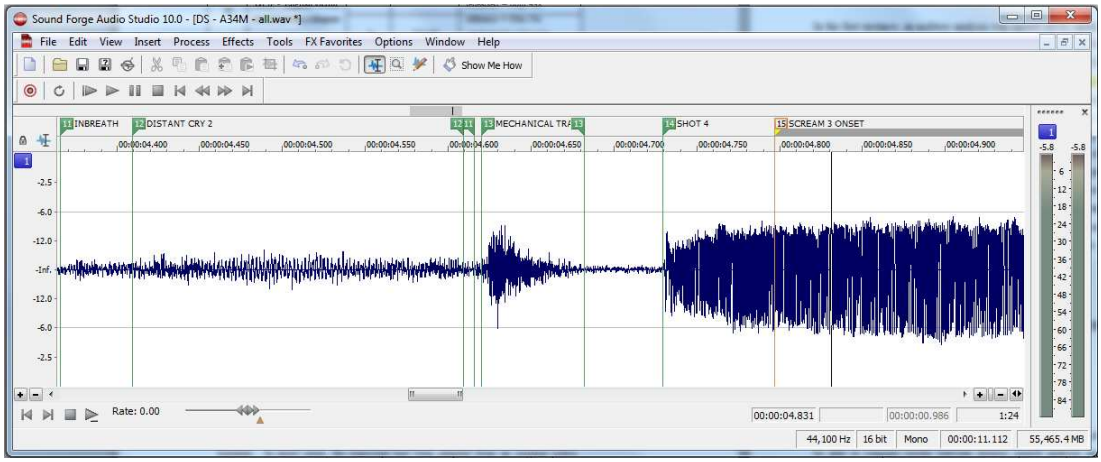
Both the acted material and digitised authentic material were edited using Sound Forge 9.0, whereby all overlaps, silences and transient noises were removed so that only speech from the speakers of interest remained. The recordings were then normalised for amplitude. For the acted material, which was recorded on two channels, only the right channel (with a microphone attenuation level set to –20dB) was selected for analysis, since the left channel was clipped during screamed sections for a few actors.

In the first instance, an auditory analysis was carried out by carefully and repeatedly listening to all authentic forensic recordings using Sony Sound Forge 9.0 in a quiet laboratory setting with closed-cup headphones (Sennheiser HD280 Pro). Markers (to indicate important one-off acoustic events) and regions (to designate important continuous events) were applied to areas of interest on the waveform and also used to isolate speech productions from the victim (Figure 3-1). These regions were then concatenated into one sound file. All edited versions were copied and saved separately from the original recording. An orthographic transcript of each authentic forensic recording was prepared in order to assist orientation with the material as well as to form the bases for the acted scripts. The transcripts drew on JPFA conventions, employing a system designed to offer the best combination of

[10] A preliminary acoustic analysis of both versions of the recording was conducted and showed that the filter did not affect acoustic measurements.

readability (for jury purposes) and detail. The transcripts resemble a script with speakers' turns alternating on lines rather than a time-aligned phonetic transcription. Unclear speech is represented in parentheses and speech with alternative interpretations is realised and marked with a forward slash. Hyphens denote incomplete or interrupted speech and ellipses show intelligible speech. Non-verbal sounds such as coughing and crying are represented in the transcript in square brackets. In most cases, the transcript had been adapted from an original police transcript. Acted material followed a pre-prepared script (described in Section 3.2.2.1).

Figure 3-1: Screen shot showing marker and region annotations to the B34M sound file using Sony Sound Forge.



3.3.2 Extracting measurements

The analysis of acoustic parameters involved extracting measurements using Praat software, versions 5.0.22 and 5.1.25 (Boersma & Weenink 2009). Where possible, only clear and continuous sections of the speakers' speech were used. However, due to the urgency of the situations, sections overlapping with the operator's speech were frequent in the authentic data set, often resulting in a smaller proportion of analysable speech. Values generated using Praat were then logged in Excel, where results could be tabulated and/or represented as graphs.

Findings from earlier studies involving naturalistic emotional data, e.g. Williams & Stevens (1972), Fuller et al. (1992), showed that features of speech that were likely to be subject to variations were identified as:

1. F0
2. intensity
3. speech tempo

These simple acoustic parameters therefore form part of this investigation in order to enable the comparison of results from distress speech analysis and other emotional speech data. It should be noted that recently, more advanced acoustic analysis techniques are being used in emotional speech analysis, e.g. inverse filtering, spectral tilt and harmonic-to-noise (H-to-N) ratios (Johnstone & Scherer 2000: 228). However, since this is the first such study of distress speech, the most basic acoustic parameters, comparable with the earliest empirical emotional speech studies, were chosen for analysis in the first instance, with a view to conducting further analysis using more recent and advanced techniques in the future, once the current investigation has been completed.

Additional acoustic variables - the vowel formants F1, F2 and F3 - were also added to the list of variables under analysis, since changes in formant frequencies have been found in speech produced with increased vocal effort, e.g. in shouted speech (Rostolland 1982), speech produced over distance (Traunmüller & Eriksson 2000), and Lombard speech, i.e. speech produced with raised vocal level against background noises (Junqua 1996).

3.3.2.1 F0

F0 mean and standard deviation (to measure the extent of high-low variation) were measured by generating a visible Praat pitch trace which was overlaid on the spectrogram. To account for possible increased F0 ranges, Praat settings were initially set at a minimum level of 75Hz and a maximum level of 600Hz for a male victim (the default), and 100Hz to 750Hz for a female (with a ceiling higher than the default), with an octave jump cost of 0.5. Where octave errors (i.e. isolated points of unrealistically high or low frequencies) occurred, the octave jump cost settings were subsequently altered until the pitch trace corresponded to pitch perception. Decreasing the values of this setting optimises Praat's ability to track rapid changes in F0 (Boersma 1993).

Problems occurred when examining screams and productions uttered at a high pitch. Despite altering the octave jump settings, the pitch trace still failed to correspond to pitch perception. In earlier reported findings from my smaller-scale MSc study using some of the same authentic data and analyses (Roberts, 2008), it was found that in a handful of examples the Praat settings had not been increased to a sufficient degree, resulting in F0 measurements recorded either an octave or a fifth below the actual level. The victims' ability to reach very high F0 had been underestimated. Greater care was taken when conducting similar analyses on the data for this investigation, and all earlier data were revisited following this error. To avoid repeat errors and to ensure accurate F0 readings across my research, the following extra steps were taken:

- pitch settings in Praat were altered and set at a minimum level of 75Hz and a maximum level of 1000Hz for a male victim, and 100Hz to 1500Hz for a female, with an octave jump cost of 0.5, though I was prepared to alter the ceilings further if necessary;
- synthetic tones from 100Hz through to 2000Hz were created in Praat to act as a computer-based tuning fork for comparison;
- difficult sections were played to other phoneticians for corroboration;
- where the voice was thought to exhibit an extreme increase in pitch, spectrogram settings were altered to display harmonics in a narrow band spectrogram and the F0 was estimated by calculating the number of harmonics in a given frequency range.

In addition, extracted sounds were examined as 'pitch objects' in Praat, whereby periodicity is analysed as a function of time using an autocorrelation-based algorithm (Boersma 1993). Sounds in this format can be manipulated so as to display all their available component harmonics (known as 'candidates') which can then be selected and deselected until the manipulated harmonic corresponds to pitch perception. Figure 3-2 is a Praat picture of a screamed production from A34M, illustrating the pitch tracking errors that can occur when using Praat's default settings. (A screen

shot of the same pitch object is shown in Figure 3-3 in order to better show the harmonic structure of the data).

Figure 3-2: Uncorrected pitch object of a distress production from Victim A (A34M) with the pitch ceiling set to the default frequency of 600 Hz.

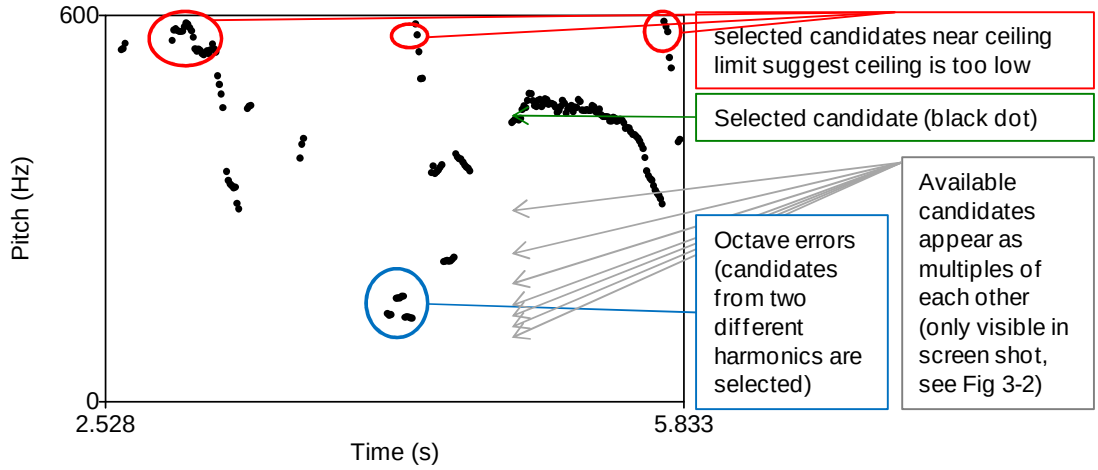
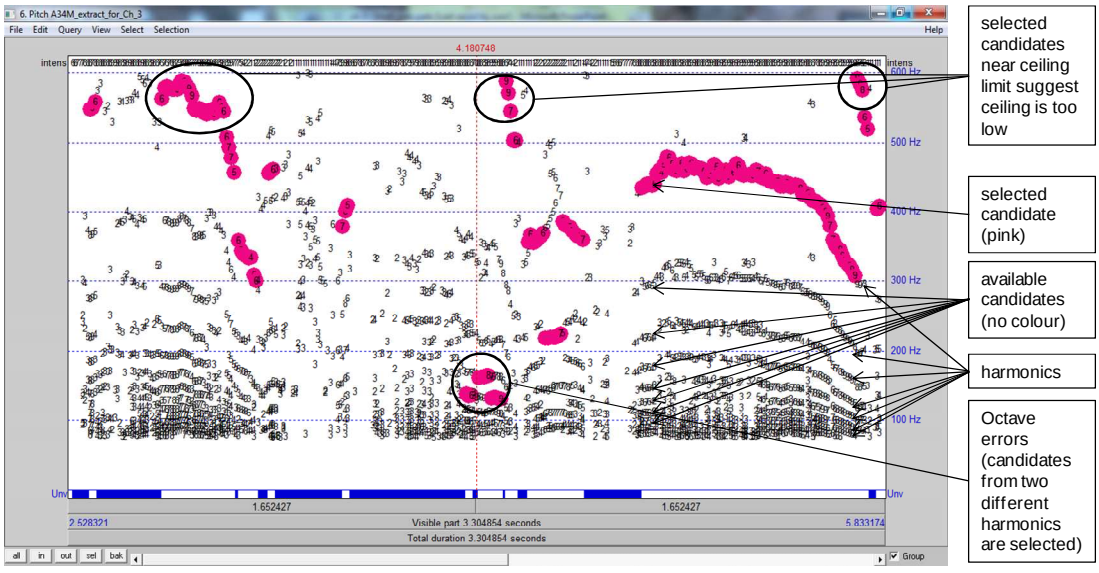


Figure 3-3: Screenshot of the uncorrected pitch object from Figure 3-2 showing greater detail of the harmonic structure and the frequencies of pitch candidates.



Raising the pitch ceiling to 1000 Hz reduces some of the tracking errors on the pitch object (Figure 3-4) but some correction by hand is also necessary in order to accurately represent the F0 in the data (Figure 3-5). For example, sections of aperiodic turbulence between A34M's screamed productions are in-breaths produced by the victim but were classified by Praat as containing periodicity (Figure 3-5). Consequently, the candidates were manually corrected to be 'unvoiced' so that their

frequencies would not be included in the acoustic analyses. Note that Figures 3-1 to 3-4 are presented using a linear Hz scale.

Figure 3-4: Uncorrected pitch object of a distress production from Victim A (A34M) with the pitch ceiling raised to 1000 Hz.

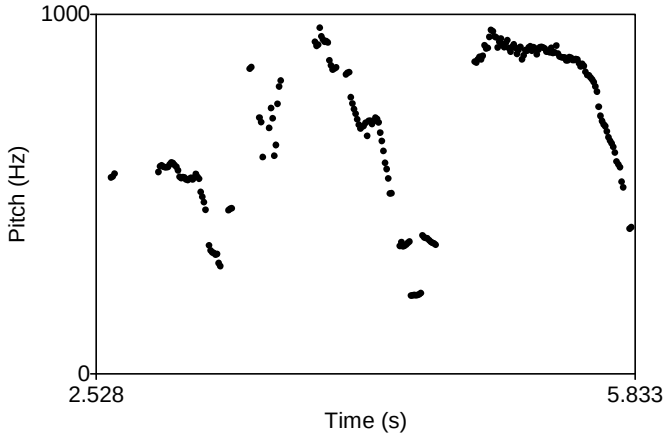
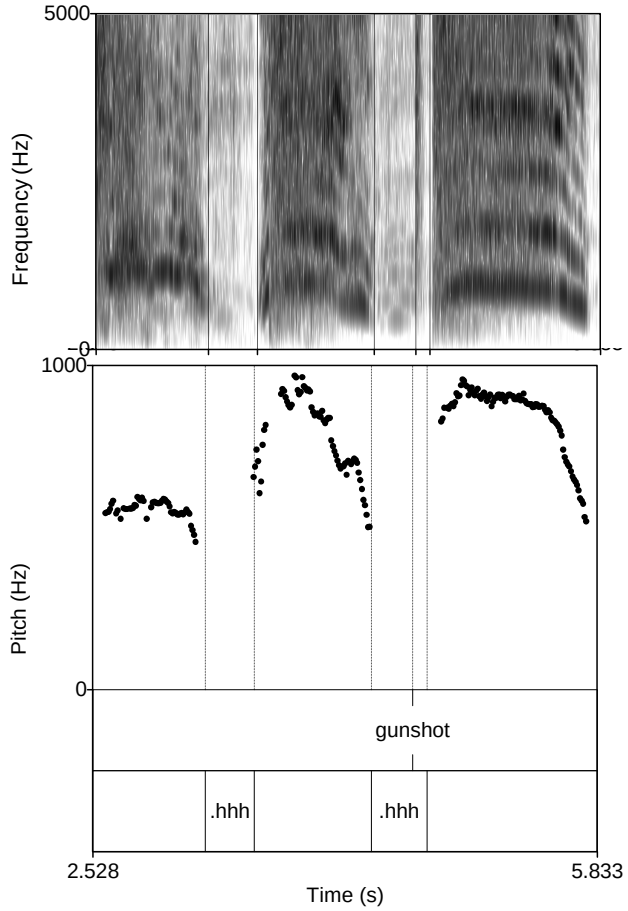


Figure 3-5: Corrected pitch object (and spectrogram) of a distress production from Victim A (A34M) with pitch ceiling set to 1000Hz.



The pitch objects play as synthesised tones at the frequency of the selected candidates and can be directly compared auditorily to the sound object. Pitch object values were recorded with the use of a Praat script, and raw mean F0 values were then converted into relative measures expressed in semitones.

Where reference data were available for authentic case material, selected excerpts typical of the victim's non-distress speech (i.e. clear and continuous with no interference or ambiguity) were amalgamated into one file per case and analysed using a Praat script developed by Philip Harrison, resulting in a Praat picture displaying the mean, standard deviation, minimum, maximum, median and alternative baseline of F0 measurements. In most forensic casework, JPFA recommend at least 90 seconds for accurate pitch analyses of (non-emotional) speech. However, due to the quantity and nature of material available in forensic casework, this amount is frequently unavailable. Indeed, in this investigation, the durations of distress material from authentic data ranged from 3 seconds (Case H) to 36 seconds (Case D).

3.3.2.2 Intensity

Intensity was measured in decibels (dB) using Praat's in-built function and measurements extracted by way of a script. The in-built default settings of a minimum of 50 dB and a maximum of 100dB were applied. Due to the uncontrolled environment and recording conditions of the victims' distress speech (i.e. no control of distance and orientation of microphone from the speaker) some cases recorded outside in open air, often with background noise, etc., I was unable to measure and quantify changes in intensity for this group to the same degree as for the actors. Since the actors were wearing head-band microphones that remained in a fixed position from their mouths, their intensity data are much more controlled, and therefore more reliable to analyse.

3.3.2.3 Tempo

For temporal characteristics, articulation rate (AR), defined as the average number of phonetic syllables per second excluding pauses over 100 ms in duration (Künzel 1997),was measured. AR has been found to have greater intra-individual stability than syllable rate (SR), which is the average number of phonetic syllables per second

including pauses (Goldman-Eisler 1968; Künzel 1997). Since speaker-specific parameters are of principal interest to the forensic speech scientist, AR was chosen over SR to be the parameter of speaking tempo. Syllables were counted and consistently quantified using a phonetic syllable count, rather than a linguistic syllable count, i.e. calculations were based on the number of syllables articulated rather than the number that would be produced if words were spoken in citation form. For example, connected speech processes, such as the elision of vowels in unstressed positions, are often found in English speech. For the word 'library', three syllables would typically be produced in citation form, but only two syllables would typically be articulated in free-flowing speech. Disfluencies, unclear sections and pauses over 100 ms in duration were edited out in line with the methodology followed by Künzel (1997:55). For consistency, the same measuring technique was applied across speech and screamed productions. However, although some screamed productions contained linguistic content, thus enabling syllables to be counted in the same way as speech productions, other screamed productions contained no linguistic content. They were perceived as loud, high-pitched vowel sounds which were often long in duration. Therefore, the AR calculation for screams with no linguistic content represents the number of screamed productions per second, rather than the number of screamed syllables per second.

3.3.2.4 Formants

Vowel formant measurements were taken for 7 monophthongal vowels: /iː ɪ ɛ a ɑː ɒ ʌ/. Wide-band spectrograms were produced in Praat for each vowel token, and vowel formant tracks, derived from Linear Predictive Coding (LPC) analyses, were generated using Praat's standard formant measuring tool. Measurements of the first three formants (F1, F2, and F3) were extracted and compiled into a log file.[11] Errors in the formant tracks were corrected by varying the LPC order and dynamic range, or they were performed by hand. The mean F1, F2 and F3 formant values were taken from averaging over a short section (typically 1-2 periods) using the maximal displacement method (Labov 1994: 165). These values were tabulated in Microsoft Excel.

[11] The formant log was created using a modified version of a script originally developed by Philip Harrison, J. P. French Associates.

Due to the lack of control over the victim's material, the number of vowel categories and tokens was not balanced across recordings. Consequently, vowel variables were not controlled across actors, since their script material was based on the uncontrolled authentic victim data. (Actors were informed that they could improvise during their recording, but they rarely deviated from the script). The vowel variables produced by each actor (or rather pair of actors) in their distress material were compared with the same vowel variables in their standardised reading passage material. Given the lack of control over the authentic distress material, actors had varying numbers of vowel tokens per vowel variable, and also varying numbers of vowel variables. Only variables that appeared in two or more cases were analysed. Table 3-3 shows the number of tokens per vowel variable per speaker.

Table 3-3: The number of vowel tokens per vowel category per speaker (d = distress speech, r = reference speech).

		Monophthongs														
		iː		ɪ		ɛ		a		ɑː		ɒ		ʌ		
Case	Speaker	d	r	d	r	d	r	d	r	d	r	d	r	d	r	Total
A	Vic A	1	1													2
	Act 1	1	6													7
	Act 2	1	5													6
B	Vic B	2	3	6	1			4	1					2	1	20
	Act 3	5	6	7	4			4	2					3	6	37
	Act 4	5	6	7	4			5	1					3	6	37
C	Vic C	1		3		2		2						4		12
	Act 5	2	6	4	4	2	8	2	3					7	6	44
	Act 6	2	6	2	4	2	8	2	3					7	4	40
D	Vic D	4		2		3		5				4		2		20
	Act 7	4	6	4	4	8	8	4	2	4	3	6	6	5	7	71
	Act 8	6	6	3	4	12	11	3		2	2	8	6	7	7	77
E	Vic E	3				1						3		1		8
	Act 9	3	5			1	10	1			3	3	7	4	5	42
	Act 10	3	5			2	7	1			1	4	7	4	4	38
F	Vic F	2														2
	Act 11	2	5													7
	Act 12	2	5													7
	Total	49	71	38	25	33	52	33	12	6	9	28	26	49	46	477

A full list of the equipment used for the acoustic analyses can be found in Appendix D1.

3.3.3 A taxonomy of distress productions

In a preliminary study on which this doctoral research is based, I acknowledged that the vocal productions of distress can be very different in nature, and that they vary impressionistically. I proposed a distress taxonomy differentiating (impressionistically) between distress speech, distress vocalisations and distress screams (Roberts 2008). I hypothesised that speech productions gradually change from speech through vocalisations to screams as distress increases, forming a distress continuum. On the one hand, this is an intuitive notion, and it has been criticised by some audiences as being obvious. On the other hand, it would be extremely difficult to measure the level of distress experienced by an individual, and therefore the continuum can not be tested. Despite our inability to validate the continuum, a tool to categorise distress productions, regardless of the level of distress experienced, would be useful when first analysing distress data in order to analyse comparable distress data. The categories expressed in the continuum - "distress speech", "distress vocalisation" and "scream" - were originally used as categories to separate distress data prior to acoustic analysis in Roberts (2008). To ensure that the taxonomy used in Roberts (2008) was reliable and replicable, a perceptual experiment was carried out before conducting the acoustic analyses of the acted and authentic data for this investigation. The data, method, findings and recommendations from this experiment are presented in the following chapter. A modified version of the taxonomy from Roberts (2008) has been put in place for the acoustic analysis conducted in this investigation. It is shown in Table 3-4. A detailed description of the taxonomy categories and a discussion of its limitations are provided in the next chapter.

The current taxonomy has evolved as an impressionistic categorisation system designed to ensure comparison of like-for-like productions across the actors and victims. For example, distressed individuals in a life-threatening situation may produce speech and/or screams. Intuitively, it would be expected that a scream would be produced with a higher F0 and amplitude than speech. As such, the findings of the acoustic analyses in Chapter 5 are presented to illustrate the differences between the four acoustic parameters not only on an aggregate level, i.e. where all the distress material from a speaker is considered in its entirety, but also as

categorised using the taxonomy, e.g. screams are compared with other screams in the data.

Table 3-4: Modified taxonomy of distress

Category	Criteria	Sub-Category	Sub-category criteria
1. *Reference speech*	intelligible, produced in a non-distressing, non-emotional context		
2. *Distress Speech*	intelligible, produced in a distressing, emotional context		
3. *Other*	unintelligible, produced in a distressing, emotional context	a.	Linguistic content
		b.	Non-linguistic content
		c.	Unclassifiable
4. *Scream*		a.	Linguistic content
		b.	Non-linguistic content
		c.	Unclassifiable

3.4 Chapter summary

This chapter re-introduced the first research question, that concerning the extent to which specific acoustic measures can be used to identify distress speech. It put forward three types of comparison to be used for investigation in this research. Firstly, authentic distress data from real victims are to be compared with non-distress data from the same individuals. Secondly, acted distress data from actors is to be compared to non-distress data from the same individuals. Thirdly, authentic distress from the victims is to be compared to acted distress from the actors. Two data sets, one of authentic forensic material and one of acted material, were presented, and the parameters under investigation (F0, intensity, AR, and vowel formant frequencies) were listed. Acoustic analysis measurements and techniques were described. Refinements to analysis techniques were also explained, and a categorisation tool to assist in comparing different types of distress response was introduced.

4. Proposing and Testing a Taxonomy of Distress

In this chapter I report on a listening experiment that was conducted before commencing the acoustic analyses of this investigation. Given that a taxonomy of distress might prove to be a useful tool in the acoustic analysis, the primary objective was to assess the reliability and replicability of a taxonomy that had originally been proposed in a preliminary study of distress (Roberts 2008). A secondary and opportunistic aim of the experiment was to test the influence of context on listeners' perceptions of distress.

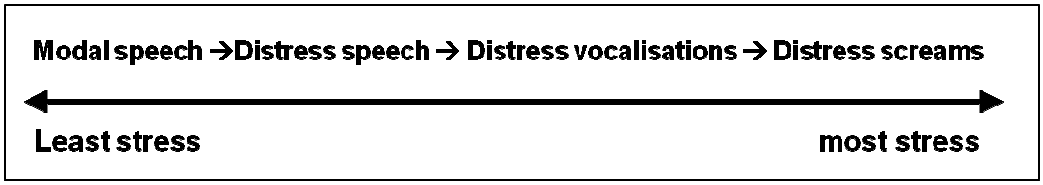
The chapter is divided into five main sections. The first part of the chapter introduces and critiques the taxonomy of distress responses as proposed by (Roberts 2008). The second part introduces the two goals of the listening experiment and describes their relevance to FSS. The third part then presents the data and methodology of the listening experiment. The fourth and fifth parts of the chapter report on findings from the listening experiment. The fourth part reports on the reliability and replicability of the taxonomy in the analysis of distress, and it proffers some amendments to the existing taxonomy before further acoustic analysis takes place. The final part of the chapter examines the effect of semantic context and background story on perceptions of distress in forensic material.

4.1 A taxonomy of distress (Roberts 2008)

The original taxonomy of distress was employed to illustrate similarities and differences in distress data collected for my MSc thesis (Roberts 2008) which acted as a preliminary study for this investigation. The taxonomy was presented in terms of a four-way categorisation system that was specifically developed to impressionistically group similar distress productions together from four victims recorded in life threatening situations (Victims A, B C and D). The speech data varied remarkably throughout these recordings, especially with regard to intelligibility and linguistic content. One hypothesis is that this variation in an individual's distress response may be attributed to the fact that the victims appear to respond to a variety of individual distress levels throughout an attack. (Roberts 2008) suggested that vocal responses form a continuum which predicts that (di)stress levels increase as the attack progresses, leading to an increase in physical and

psychological (di)stress (presumably due to pain and fear), a decline of intelligibility, and a reduction of perceptible linguistic content, as illustrated in Figure 4-1.

Figure 4-1: The distress continuum (Roberts, 2008)



The original classification proposed a continuum divided into four categories, as outlined below:

1. reference/modal speech (representing non-emotionally aroused speech, such as control speech if alternative data were available for the victim from a non-distressful situation);[12]
2. distress speech (emotionally aroused, intelligible productions with linguistic content);
3. distress vocalisations (unintelligible productions – it is unclear if linguistic content is present or not);
4. distress screams (emotionally aroused productions with no linguistic content).

Given the limited material available in some of the recordings, productions from victims progressing through the entire continuum were not encountered, i.e. no recording allowed for all four categories to be represented by the one victim. Most victims' productions were classified using two or three categories.

[12] The term 'modal speech' was originally used in Roberts (2008) and during the listening experiment described in this chapter. It did not refer to manner of phonation but to everyday speech produced in non-violent, non-distressful conditions. It was not proposed that everyday speech is without emotional arousal but it was suggested that it is distinct from the type of speech as produced following, during, or immediately after a violent attack. To avoid confusion, and following post-experiment revisions to the taxonomy, 'reference speech' or 'non-distress' replace the 'modal speech' category for the remaining investigation.

4.1.1 Justification of the taxonomy

The primary motivation for categorising distress productions was to be able to compare like with like. For example, Victim A produces discernible speech and screams, some of which reach 968 Hz in F0, yet Victim B produces no screams in his distress material. Victim's B's speech, although produced with an increased F0 (maximum F0 is 307 Hz), is not as increased as Victim A with respect to F0, nor is it as variable (standard deviation in semitones is 3.9 for Victim B versus 10.7 for Victim A). Clearly, a binary distinction between reference speech and distress speech is too crude to allow us to compare the finer nuances of variation in distress productions within and across speakers. Secondly, the research presented here originated as an exploratory study of distress productions, describing the different types of vocal responses exhibited by the victims (and later in actors). To my knowledge, this has not been performed in previous studies and so the production of a categorisation protocol may in itself be revealing and worthy of interest. It would allow the user to consistently quantify impressionistic differences between speech productions.

It was hypothesised that the victims' productions occupying the extremes of the taxonomy would be less prone to dispute, i.e. speech produced in everyday, non-violent situations would be easily recognised as being non-distress speech, and full-bodied screams would be recognised and easily differentiated from the other taxonomy categories. The boundaries of categories mid-continuum are not as clearly delineated. It often difficult to identify the communicative intention of the victim during his/her productions and so the classification of the production may be questionable. For example, Victim G suffers head injuries and can be heard vocalising (see §3.2.1.7 for more detailed case information). It is unclear if she is wailing, moaning or attempting to produce speech. Additionally, throughout the course of an utterance, a victim may start to vocalise in a manner that switches between categories. Victim D (§3.2.1.4), for instance, produces an alveolar nasal [n] with high F0, followed by a long, open vocoid with increasing F0 peaking at 1400 Hz. Possible interpretations of this production include a scream carrying linguistic content, a vocalisation with questionable linguistic content, or speech. The word 'no' would be appropriate in this context (she is trying to escape from a man who has just

shot her friend) but based on the material given, it cannot be determined with certainty what her intended production was.

4.1.2 Criticism of the original taxonomy

It should be acknowledged that there is no way to validate distress categorisations as presented in the original taxonomy since there is no way to obtain 'ground truth', i.e. it is not possible to measure the level of distress as experienced by the victim or to correlate distress level with responses. However, a categorisation system that allows listeners to impressionistically group distress responses together would still be of value to those researching distress, especially in the preliminary stages of investigation. No other distress categorisation schemes have been reported, and so a listening experiment was carried out in order to assess if the taxonomy presented in Roberts (2008), despite its theoretical flaws, might be a reliable and replicable categorisation system for the current investigation. The listening test specifically aimed to assess the robustness and applicability of the taxonomy by other forensic practitioners before the analysis of both acted and authentic datasets for this investigation (in terms of the parameters described in chapter 3) was conducted.

4.2 Listening Experiment Research Questions

The listening experiment was developed in order to address concerns about the implementation of a distress taxonomy as a way of comparing distress productions across speakers and listeners (§4.1.2) as well as to assess the influence of background and contextual information on listeners' perceptions of distress (§4.2.2)

The experiment seeks to address two principal research questions:

1. Can the original taxonomy be easily and consistently applied by other forensic practitioners using the same forensic material?
2. To what extent and in what ways does contextual (semantic and background) knowledge of the material affect the listener's perception of the extract?

4.2.1 The reliability and replicability of the distress taxonomy

In addressing this first question, the study aims to investigate whether the existing classification system is a reliable and replicable tool for listeners such as forensic practitioners who may be called upon to analyse distress responses. As with all practical taxonomies, there will always be indeterminacy and arbitrary decisions (Sinclair & Coulthard 1975: 16); however, a high level of agreement would be anticipated amongst a group of similarly trained forensic practitioners.

4.2.2 The influence of context on listeners' perceptions of distress

For this secondary question, the study aims to explore the effect of higher-order contextual information - semantic and background story - on interpretations of forensic material, namely the listeners' perception of distress. Although the principal aim of the listening experiment is to test the reliability and replicability of the original distress taxonomy using extracts from authentic material, I took the opportunity to include an additional variable in the experiment design by playing some extracts with which background information and context were available to the listener, and some for which this was not provided. The reason behind adding this additional layer of complexity to the experiment is that forensic practitioners are often asked to transcribe and interpret recordings that are brief and of poor quality. To avoid preconceptions on the part of the practitioners biasing the transcription, a 'bottom up' approach is encouraged (Fraser 2003), whereby the listener first undertakes the task without contextual information, working only from the recorded sound. This entails that the forensic practitioner undertake the task 'blind', i.e. without contextual information, a technique which is therefore viewed in a court of law as a more objective method of transcription than where the listener transcribes using a 'top down' approach. On the one hand, some practitioners have conceded that their work has been impeded by the lack of contextual information (Hirson & Howard 1994). Others, however, have written reports that may give the impression that that the significance or linguistic content of an unclear utterance was resolved mainly or wholly from examining its internal phonetic and acoustic properties (French 1990; Rose 2009).

4.3 Experiment design

4.3.1 Stimuli

Twenty four utterances drawn from authentic forensic recordings were presented as audio stimuli. The recordings contained distress productions by victims from eight criminal investigations (victims from cases A-H) where an attack had been imminent, in progress, or recently completed. Information concerning the circumstances of each case is provided in §3.2.1. All victims originated from either England or Wales and spoke English as their native language. Stimulus materials were, in general, balanced in terms of including the voices of victims of different sexes and ages. Each of the eight recordings generated between two and six extracted productions to capture the variability in the speech of the victims throughout the recording. Given the nature of these recordings, there was little opportunity to represent the non-distress speech category, though non-distressed material was available from other recordings for two of the eight victims (Victims A and B).

The content of the victims' productions was often related to ongoing circumstances, e.g. injuries or reference to locations; pleading with the attacker was also frequent. Where possible, utterances were chosen that could be interpreted neutrally, such as utterances containing locations or words such as 'boyfriend' and 'court'. These words did not presuppose a violent or distressing scenario. However, several utterances clearly remained non-neutral, e.g. the words 'shot', 'stabbed', and screamed productions could be heard.

The stimulus materials were extracted from the digitised versions (44.1kHz, 16 bit) of the original recordings using Sony Sound Forge. Where possible, material overlapping with speech from emergency services operators and other agents (such as witnesses and/or attacker(s) at the scene) was avoided.

Each of the twenty four extracted sound files was played in two conditions:

i) 'without context' - the production was heard in isolation, devoid of any speech information either preceding or following the extract, and no information surrounding the circumstances was given;

ii) 'with context' – the production was heard with preceding and/or following semantic context (in some cases including previous/following turns of the emergency services operator), and background information concerning the circumstances of the recorded attack was provided.

Twelve of the extracts were labelled 'Pool α' and the remaining twelve as 'Pool β' (each pool was as balanced as possible in terms of having an equal mix of male and female voices, ages, and representations of the four categories). A further eight extracts, devoid of surrounding context and background information, were taken from the same corpus of authentic forensic recordings to act as controls. These extracts were only played in one condition - 'without context' - and were repeated later in the experiment to test whether participants' behaviour changed following increased exposure to the experiment.

4.3.2 Participants

Sixteen members of the forensics research group in the Department of Language and Linguistic Science, University of York, participated in the experiment. All participants had either taken or taught the postgraduate 'Introduction to Forensic Speech Science' and 'Research in Forensic Speech Science' modules offered as part of the MSc in Forensic Speech Science during the academic years 2007-2010.

4.3.2.1 The 'inexperienced' listeners

The inexperienced group was comprised of ten postgraduate student phoneticians (three male, seven female) who had all had some experience of listening to authentic case material (though not of the nature presented in the experiment) during their Masters studies. They were aged between 22 and 25 years.

4.3.2.2 The 'experienced' listeners

The experienced group was comprised of six participants (four male, two female) who had extensive forensic casework experience. All had acted as expert witnesses. The ages of the experienced group ranged from 27 to 57 years.

Most participants were native English speakers originating from England and Scotland, though three female student phoneticians had other first languages (two speakers of German, one of Thai) and two student phoneticians spoke English as a

native language from areas other than the UK (one female from Canada, one male from the United States).

Members of the forensics research group were deliberately chosen to act as participants due to their willingness to listen to potentially distressing material, their readiness to participate in forensic research and their familiarity with the poor quality of authentic forensic recordings (which is typically more degraded than material used by non-forensic practitioners).

4.3.3 Procedure

Prior to taking part in the experiment, participants were asked to read an experiment information sheet which also summarised the categorisation system used in the distress taxonomy (Appendix E1). They were invited to raise any questions or concerns before agreeing to take part and signing the consent form (Appendix E2). The experiment involved listening to extracts from six 'blocks'. The first and fourth blocks contained the eight randomly-ordered control extracts (all of which were devoid of semantic context and in relation to which no background information was provided), whereas the remaining blocks each contained twelve randomly-ordered extracts. Two of these blocks contained the 'with context' versions of the extracts, whereas the other two blocks were comprised of 'without context' extracts. An example of the visual stimuli provided as part of the listening experiment when assessing audio extracts with and without contextual information is given in Appendix E3. To minimise memory effects, a block of extracts which had been played in one condition would never follow or precede a block of the same extracts in the other condition, and each block was always quasi-randomly ordered. As illustrated in Table 4-1 and Table 4-2 below, two experiments were run, varying the order of presentation to control for a possible order effect.

Table 4-1: Experiment A design of the listening test

Block A	Block B	Block C	Block D	Block E	Block F
Control (8 extracts)	Pool α with context (12 extracts)	Pool β without context (12 extracts)	Control (8 extracts)	Pool β with context (12 extracts)	Pool α without context (12 extracts)

Table 4-2: Experiment B design of the listening test

Block A	Block B	Block C	Block D	Block E	Block F
Control (8 extracts)	Pool α without context (12 extracts)	Pool β with context (12 extracts)	Control (8 extracts)	Pool α with context (12 extracts)	Pool β without context (12 extracts)

The experiment was delivered via PowerPoint Presentation 2007, and audio files were played through closed-cup Sanako Tandberg Educational headphones (headset model SLH-07).

Each extract lasted no more than a few seconds (in the 'without context' condition sometimes even less) and was played three times with three seconds of silence inserted between repetitions. For each extract, each participant was asked to:

(a) categorise each extract in terms of the four-way classification corresponding to the degree of distress s/he perceived the stimulus to represent;
(b) rate the extract on a 5-point scale specifying the degree to which s/he perceived the stimulus material to have linguistic content.

Participants were asked to evaluate the extracts on these two scales in order to explore the relationship between the features of emotional arousal, intelligibility and linguistic content which formed the basis of the original classification system. When categorising an extract for perceptible distress, participants were asked to arrive at their decisions using the classification system described in their information sheet (Appendix E1). Although some utterances were clearly not neutral in respect of content, participants were asked to consider the manner in which the utterances were produced, rather than what was said, when categorising distress using the four-way classification. The perception of linguistic content was to be rated on a separate five-point Likert scale, with 1 being 'clear linguistic content' and 5 'no apparent linguistic content'. When rating for linguistic content, both the manner in which the utterance was produced and the content of the utterance were to be considered.

Space was also provided on the response sheet to encourage participants to make notes and/or offer transcriptions (orthographic and/or IPA). A sample response sheet can be found in Appendix E4. Only a few experienced listeners volunteered transcriptions, but these were inconsistent in manner and distribution. These qualitative data do not feature as a main point of investigation in the current experiment. Participants were able to progress through the six blocks at their own speed but were not allowed to repeat sound files or to return to previous extracts to change their responses.

4.4 Listening Experiment Results I - reliability and replicability of the distress taxonomy

To assess the reliability and replicability of the original classification system, the variability of extracts and participants will first be considered. Findings concerning the level of agreement between participants' categorisations of extracts and the extracts' original classification by the author will then be presented to assess the reliability and replicability of the taxonomy.

The following results sections will only report on participants' responses from extracts in non-control blocks of the experiment (Table 4-1 and Table 4-2). The responses from the two control blocks (blocks A and D of the experiment) were analysed to test whether increased exposure of the experiment resulted in changes in participants' responses, but a between-subjects one-way ANOVA did not show a significant effect for changes in either categorisation or rating of extracts across participants' control responses. This means that the participants were consistent in their responses to extracts from the control blocks, showing that exposure to the experiment did not alter their perceptions. For this reason, the responses from the control block extracts are not included in the results sections that follow.

4.4.1 Variability of extracts

Listeners showed varying degrees of consensus when scoring each extract. Figure 4-2 and Figure 4-3 show the listeners' mean and standard deviation of distress categorisation scores for each extract heard with and without contextual information (regardless of listener experience). Figure 4-5 and Figure 4-6 show the same information but for linguistic content scores.

Figure 4-2: Mean distress categorisation scores of extracts in both 'with/without context' conditions.

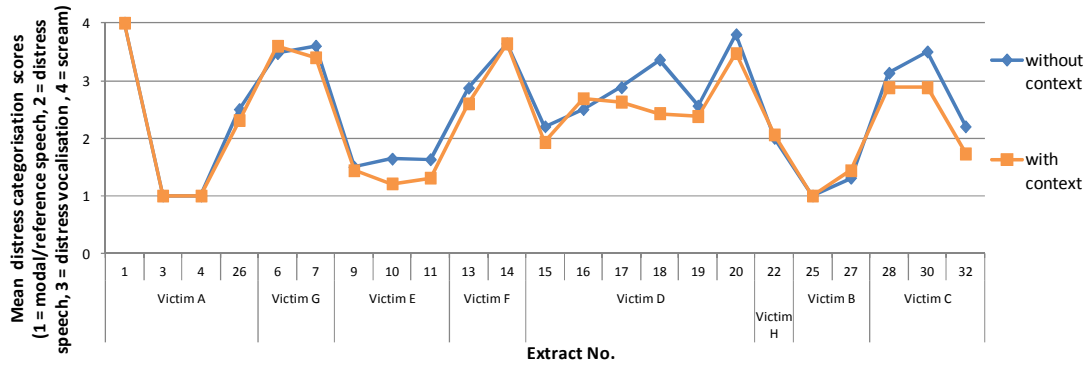
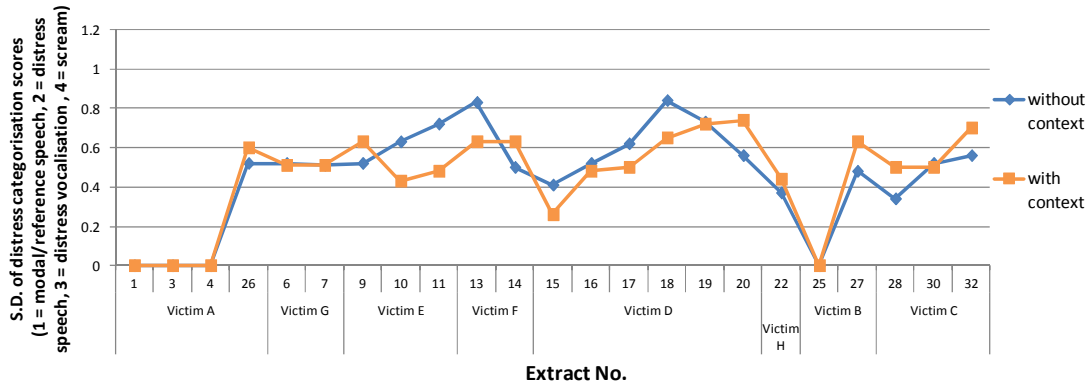


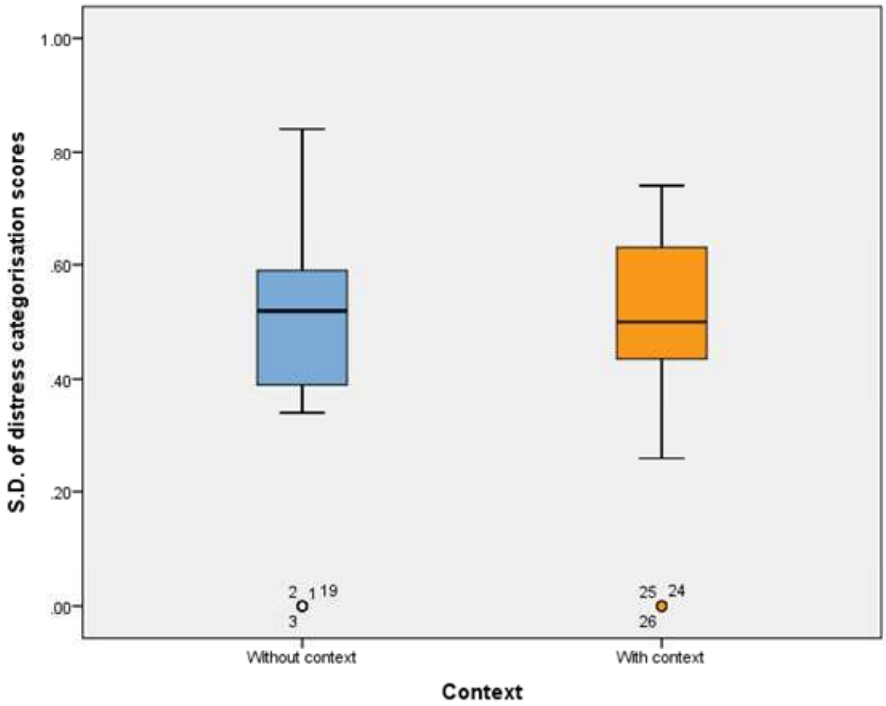
Figure 4-3: Standard deviation of distress categorisation scores of extracts in both 'with/without context' conditions.



Although the majority of extracts' scores demonstrate disagreement between participants with respect to their distress category ratings, four of the extracts (1, 3, 4 and 25) received a unanimous categorisation. Three of these extracts were from the same recording (1, 3, and 4) and three were non-distressed utterances from the victims (3, 4, 25). Common to all four is that they have been categorised as occupying the extremes of the distress categorisation continuum, receiving the same score from all listeners in both conditions. Disagreement mainly occurs for extracts occupying the middle of the cline where category boundaries are less well-defined. The mean and standard deviation of the distress categorisation scores across extracts are, on the whole, higher in the 'without context' condition – the average mean across all extracts falls from 2.47 to 2.29 and the average standard deviation drops from 1.08 to 1.04 – demonstrating that extracts heard without context are perceived to contain more distress and are harder to categorise. A paired-samples t-test was conducted to compare distress category rating when samples were heard in 'without

context' and 'with context' conditions. There was a significant difference in the means of distress categorisation scores for extracts heard without context (M = 2.49, S.D. = 0.97) and those heard with context (M = 2.31, S.D. = 0.94) conditions; t (22) = 3.36, p = 0.03. Due to the non-normal distribution of the distress category standard deviation scores, these data were not subject to statistical testing. (An F-test, for example, would be sensitive to the non-normality and would give a misleading result). Therefore, the standard deviation scores are compared graphically below using boxplots (Figure 4-4). The boxplots suggest differences between the two conditions - the variance of extracts' ratings heard without contextual information has lower median and upper/lower quartiles - and reveal outliers in the form of extracts which received unanimous or similar ratings. There is a tendency for extracts to receive more consistent distress categorisation ratings when the extracts are heard with contextual information.

Figure 4-4: Extracts' distress category standard deviation scores.



There are, however, some cases where the opposite of these trends is found. In extract 27, the presence of context leads to an increase in the mean and standard deviation of perceived distress level from 1.31 to 1.44 and 0.48 to 0.63 respectively. The extract concerns a 42-year-old man giving the emergency services operator his

location after he has been stabbed in the stomach. The production contains 3 monosyllabic words detailing the name of the street. The production is not clearly articulated, and where participants had volunteered transcriptions, there were several different versions. In most cases, the first two words were parsed as being a bisyllabic word, which might be reasonable given that many place names contain a bisyllabic word followed by a monosyllabic one, e.g. Station Road, London Road, Manor Lane etc. The presence of context here appears to have made the listener more sensitive to an attempt by the victim to answer the operator with a place name. The victim's inability to properly articulate this information may have also provided the listener with an indication that the victim has sustained some serious injury. Together, these factors may have led to a high perceived level of distress.

Similarly, in extract 32, where a 47-year-old man has telephoned the emergency services to report that he has been stabbed in the throat, the standard deviation of the distress categorisation scores increases with context from 0.56 to 0.70. The utterance in this extract, “I’m dead”, is delivered within the typical male F0 range, is intelligible, and easily transcribed. Despite the circumstances, little distress is perceived within this extract and so when context is provided, the listeners are more aware of the victim’s injures and rate the extract as containing more distress.

One of the most difficult extracts to classify is extract 18, concerning a 15-year-old victim producing a voiced alveolar plosive [d] followed by a diphthongal vocoid with a open-mid, front, unrounded nucleus and a close, central off-glide [ɛɻ̟] at a high F0. This extract has the highest standard deviation of distress categorisation score (0.84) when heard without context. At least two interpretations are available given the context of the utterance, but due to the extremely high F0 and inability to parse the victim’s speech attempts, categorising this extract remains ambiguous.

Equally, extract 13, receiving a similarly high standard deviation of 0.83 when heard without context, is from a 27-year-old woman calling the emergency services after having been shot in the neck and lower jaw, and contains a production at high F0. The production could be interpreted as a voiceless bilabial plosive and lateral approximant cluster [pl], followed by a close frnt vowel [i] and some frication [i^{h}].

The production could be in answer to the operator's previous question, "Emergency, which service?" with either 'police' or 'please' being a plausible response – the latter is possible if we consider the victim might be intending to say the 'please help' or 'please hurry' (the victim is in urgent need of medical assistance and the attacker is still downstairs). However, due to her injuries, the woman has difficulty articulating and so when the extract is heard without context, listeners vary in their willingness to assign intelligibility.

Figure 4-5: Mean linguistic content scores of extracts in both 'with/without context' conditions.

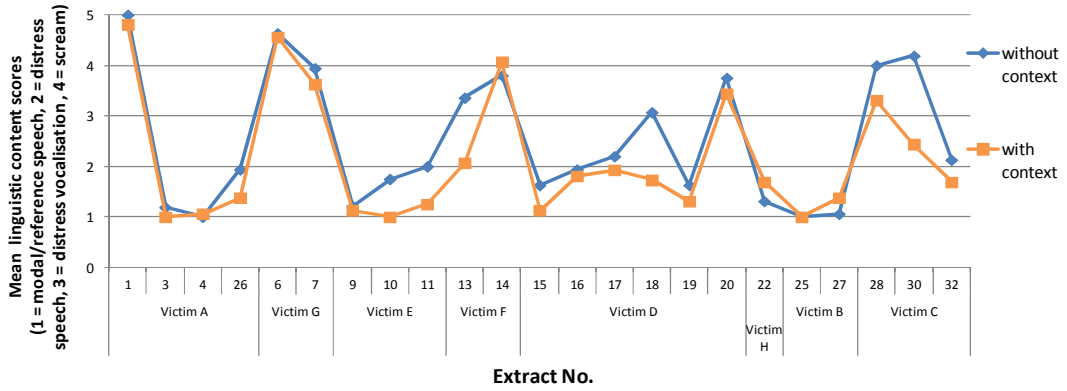
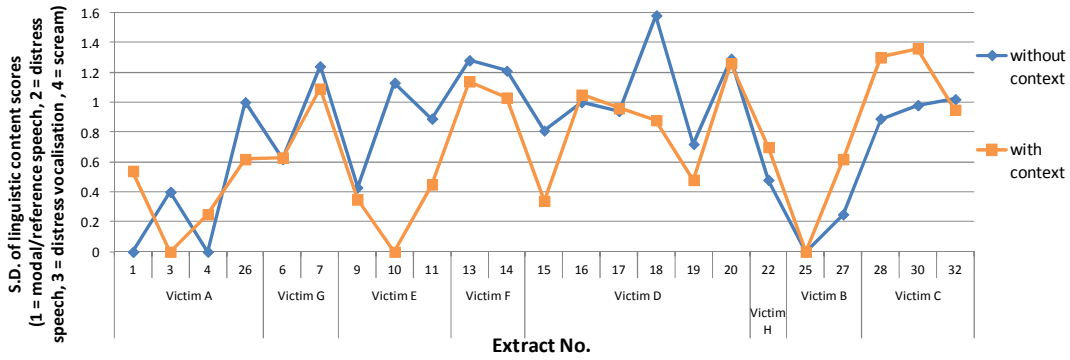


Figure 4-6: Standard deviation of linguistic content scores of extracts in both 'with/without context' conditions.

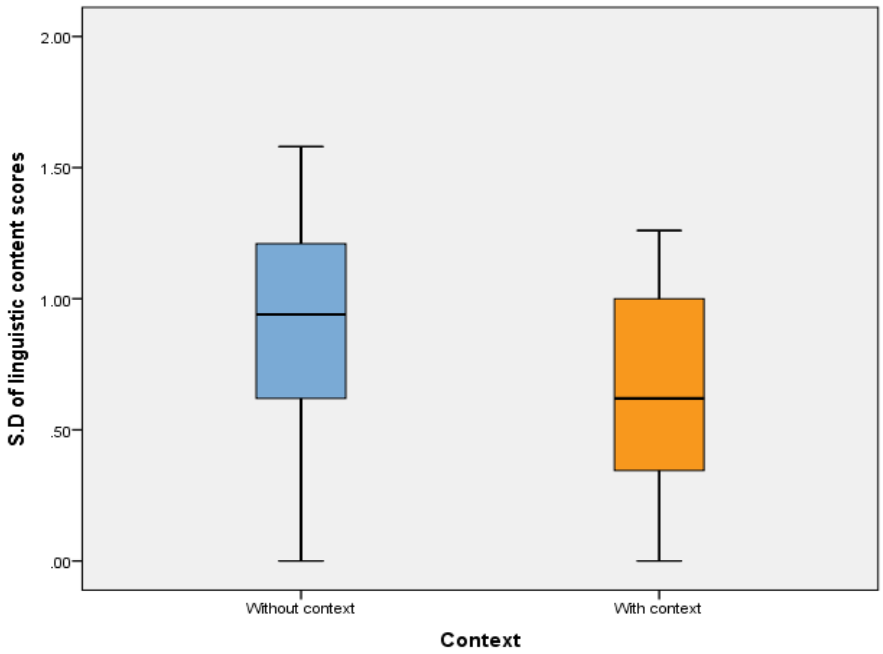


As expected, extracts rated for distress categorisation with the highest and lowest degrees of consensus by participants maintained their status quo in ratings of linguistic content. Extract 25 retains its unanimous scoring by all participants in both conditions, whereas extracts 1 and 4 retain unanimous ratings in the 'without context' condition, as does extract 3 in the 'with context' condition. All three have low standard deviations in the opposite condition, i.e. extracts 1 and 4 have low standard deviations in the 'with context condition' and extract 3 has a low standard

deviation ‘with context’ (0.54, 0.25 and 0.4, respectively). The ratings for these extracts signify a high degree of consensus among most, but not all, participants. Extracts 18 and 13 continue to have high standard deviations in linguistic ratings as well as distress categorisations for samples without context (1.58 and 1.28 respectively), thus indicating a continued lack of consensus across participants, though the presence of context reduces the standard deviation for extract 18 (from 1.58 to 0.88), i.e. hearing the extract with context improves the interpretation of the production for most listeners.

Akin to the patterns for the distress production categorisations, means and standard deviations generally tend to be higher in the ‘without context’ condition when rating linguistic content, indicating a decrease in perceptible linguistic content and a greater level of disagreement between participants’ responses when extracts are played devoid of contextual information. The average mean across all extracts falls from 2.52 to 2.12 and the average standard deviation drops from 1.54 to 1.43 when the extracts are heard with context. A paired-samples t-test to compare mean linguistic content ratings when heard with and without contextual information revealed that there was a significant difference in the scores for extracts heard without context (M = 2.51, S.D. = 1.29) and those heard with context (M = 2.12, S.D. = 1.22) conditions; t (22) = 3.49, p = 0.02. There was a non-significant F-test result for standard deviation linguistic content score, possibly due to the low number of observations, but boxplots (Figure 4-7) show a tendency for extracts heard with context to have lower median and quartile scores, suggesting higher cross-listener agreement when extracts are heard with contextual information.

Figure 4-7: Extracts' linguistic content standard deviation scores.



Extract 10 highlights the influence of context on perception, showing a unanimous decision by all listeners to rate the production as ‘non-distress speech’ (i.e. a rating of 1) in the ‘with context’ condition, but a divided score by listeners when the context is absent. This extract contains the words ‘been shot’ spoken by a young woman calling the emergency services after having been shot in the chest with an air gun. The production was clearly articulated and delivered in a typical female F0 range. Without context, participants varied in how much distress they perceived, with some opting for non-distress speech and others choosing distress speech (mean score is 1.75 and standard deviation is 1.13).

Extract 30 is noteworthy since although there was some variation in distress categorisation scores, scores for linguistic content rating were much more varied. In the ‘without context’ condition, the mean score was 4.19, i.e. the extract was deemed to be strongly lacking in perceptible linguistic content, with a standard deviation of 0.98. In the ‘with context’ condition, the extract had a mean score of 2.44, i.e. it was perceived to have some linguistic content, and a standard deviation of 1.36. This extract is from the recordings of a 47-year-old man who had been stabbed through the throat. Participants consistently recognise this production as a vocalisation rather than distress speech, but vary in their scores when assigning perceptible linguistic content. The production sounds as if it could contain linguistic information in the

form of a palatal approximant [j] followed by (mid-)open front vowel [ɛ̝], though it could also be the victim vocalising in pain or fear. Both are equally plausible, since it is unknown whether the production is an affirmative response to the operator's question, "Is it for yourself?" or as a response to external events. When participants hear the production with context, i.e. with the operator's question preceding it, their scores reflect a higher degree of perceived linguistic content.

4.4.1.1 Summary of individual extracts

Individual extracts elicit mixed degrees of cross-listener agreement. The standard deviation for each extract (as rated by all participants) varied from 0 to 0.84 for distress category, and from 0 to 1.58 for linguistic content. Unanimous scores with a low standard deviation tended to occur when listeners were rating productions originally classified by the author as occupying the peripheries of the taxonomy (non-distress speech and scream). Where standard deviations were high, that signified multiple interpretations of the production in question. The lower standard deviation scores in the 'with context' condition imply that the presence of context, in most cases, aided the listeners by narrowing the possible interpretations of the production, thus leading to more agreement between subjects.

In addition, when heard with context, extracts yielded statistically significantly increased mean scores, suggesting that the participants perceived the extracts to belong to a less distressed category of the taxonomy and to contain more linguistic content.

4.4.2 Individual performances

Performances by individual participants varied substantially throughout the experiment. Figure 4-8 and Figure 4-9 show the mean and standard deviation of each participant's distress categorisation score, regardless of the listener's level of experience, across all extracts heard with and without contextual information.

Figure 4-8: Mean distress categorisation scores by participants in both 'with/without context' conditions.

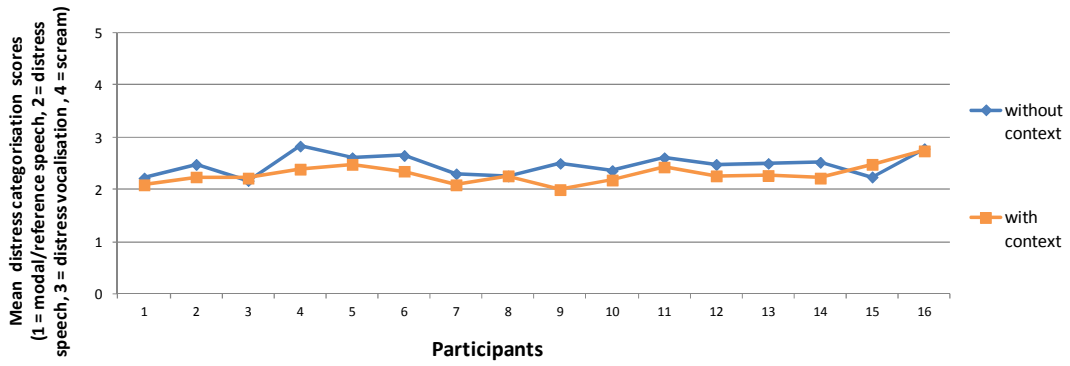
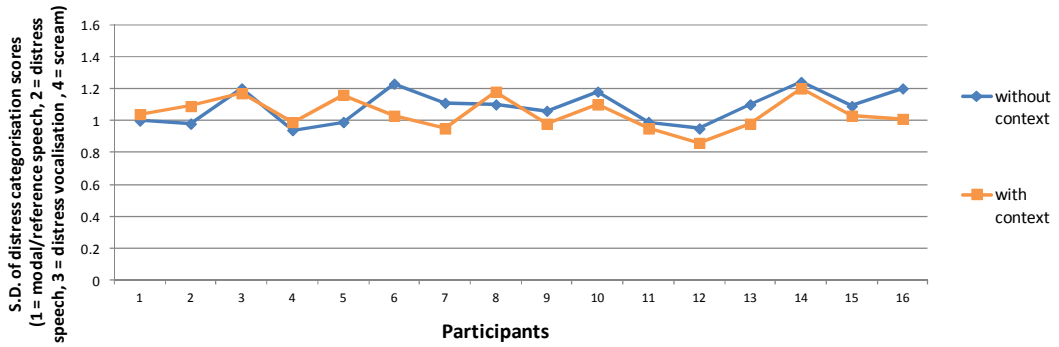


Figure 4-9: Standard deviation of categorisation scores by participants in both 'with/without context' conditions.

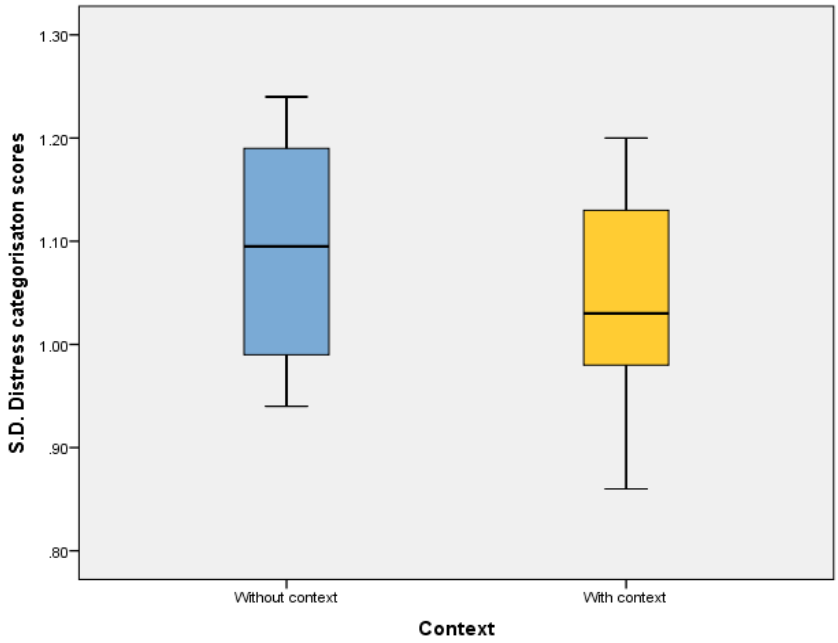


In the 'without context' condition, it can be seen that participants 16 and 3 have the same standard deviation (1.20) but different mean scores (2.17 vs. 2.78 respectively). Participant 16 appears to consistently categorise the distress productions as more distressed than does participant 3. On the other hand, participants 12 and 14 have similar mean linguistic rating scores in both conditions (2.48 and 2.26 for participant 12, and 2.52 and 2.22 for participant 14), but very different standard deviations (0.95 and 0.86 for participant 12, and 1.24 and 1.20 for participant 14). This suggests that participant 14 scores the extracts using values from the entire four-point scale, whereas participant 12 may be prone to categorising distress productions using only mid-values on the scale.

A paired-samples t-test was conducted to compare individuals' distress categorisation ratings when heard in the 'without context' and 'with context' conditions. There was a significant difference in the scores for listeners when extracts were without context (M = 2.47, S.D. = 0.20) and those heard with context

(M = 2.29, S.D. = 0.18); $t(15) = 3.87$, $p = 0.02$. There was a non-significant F-test result for standard deviation distress categorisation scores, which might be due to the low number of observations, but the boxplots in Figure 4-10 illustrate that extracts heard with contextual information tend to have lower median and quartile scores, suggesting reduced variation in participants' distress categorisation ratings.

Figure 4-10: Participants' distress categorisation standard deviation scores.



Turning to Figure 4-11 and Figure 4-12, which show participants' ratings of linguistic content, participant 16 once again has the highest mean (3.22 in both conditions) and scores markedly higher than some other participants, especially participant 4 (means of 2.35 and 2.0 in 'without context' and 'with context' conditions respectively), if their similar standard deviations are taken into account (1.59 for participant 16 and 1.58 for participant 4 in the 'without context' condition). Participants 5 and 15 have similar mean scores (1.52 and 1.65 respectively) but different standard deviations in the 'with context' condition (0.95 vs. 1.19), showing that participant 5 is more likely to consistently rate extracts lower than is participant 15.

Figure 4-11: Mean linguistic content scores by participants in both 'with/without context' conditions.

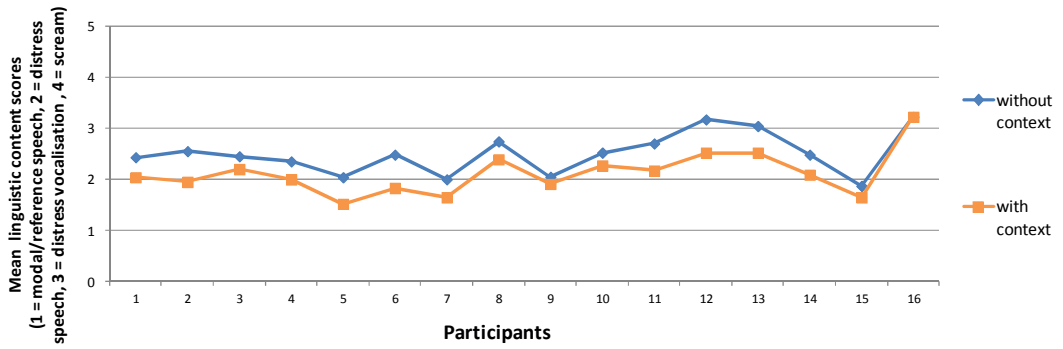
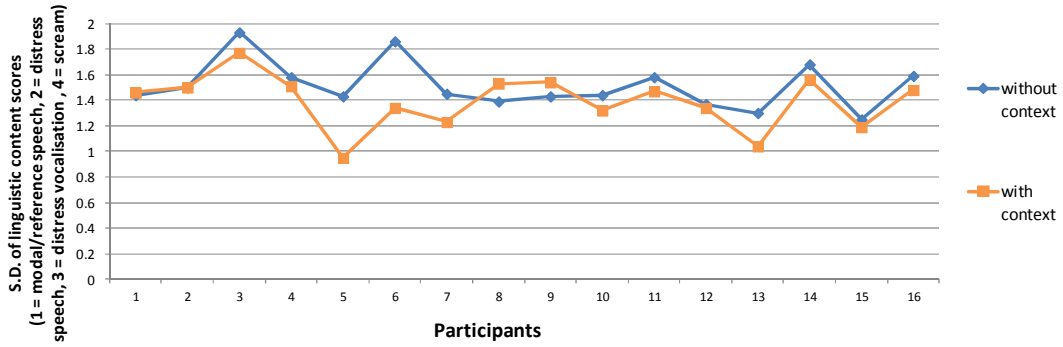
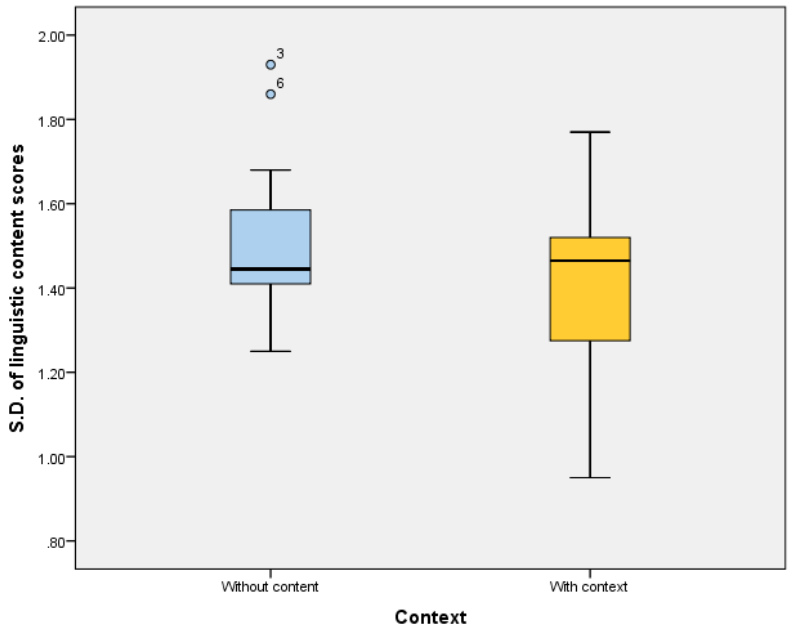


Figure 4-12: Standard deviation of linguistic content scores by participants in both 'with/without context' conditions.



A paired-samples t-test comparing participants' linguistic content ratings when samples were heard with and without contextual information revealed that there was a significant difference in the scores from individuals when extracts were heard 'without context' (M = 2.51, S.D. = 0.40) versus 'with context' (M = 2.12, S.D. = 0.42); $t(15) = 8.27$, $p < 0.001$. Akin to the standard deviation of distress categorisation scores, there was a non-significant F-test result for the standard deviation of linguistic content scores. However, Figure 4-13 shows that extracts heard with contextual information tend to have a higher median score and larger interquartile range. It suggests reduced variation in participants' distress categorisation ratings, and reveals two outliers representing participants with the highest standard deviation scores in the 'without context' condition.

Figure 4-13: Participants' linguistic content standard deviation scores.



The average mean and standard deviation of participants' distress categorisations and linguistic content ratings across all extracts once again highlight the trend that the presence of context reduces the level of perceived distress, increases the level of perceived linguistic content, and reduces the amount of variation in participants' ratings, i.e. context either inhibits participants from using all of the scale (scores are clustered in the middle of the scale for all extracts), or it causes participants to narrow their choices of score for each extract (their scores make use all of the scale as before but are concentrated around a certain point for each extract, forming multiple clusters). Given that the extracts were deliberately chosen to test the breadth of the two scales, the latter option may be more plausible. The extent of this trend varies across individuals. For instance, the presence of context markedly reduces participant 55's standard deviation of linguistic content ratings (from 1.43 to 0.95), whereas participant 2's standard deviations do not change at all (1.5). Similarly, participant 6's standard deviations of distress production categorisations decrease with context (from 1.23 to 1.03), yet participant 11's standard deviation barely changes over the two conditions (0.99 to 0.95).

4.4.2.1 Summary of individual performances

Performances by individuals varied. The participants' mean distress category score (calculated by averaging over all extracts) varied from 2 to 2.83, with standard deviations ranging from 0.86 to 1.24. For linguistic content ratings, the mean scores

spanned from 1.52 to 3.22, with standard deviations ranging from 0.95 to 1.93. Some participants rated extracts consistently higher or lower than did their counterparts; others had similar scores due either to cross-participant agreement, or because of consistently rating in one area of the scales, or rating across the whole scale and averaging in the same area as other participants. In accordance with findings from the average scores of individual extracts in §4.4.1, ratings by individual participants increased for both distress category and linguistic content when heard with context, reiterating the view that the participants perceived the extracts to belong to a less distressed category of the taxonomy and to contain more linguistic content when semantic context and case background is provided.

4.4.3 Group performances

To examine the effect of experience on the participants' performances, a by-subjects analysis was conducted to see if the group of experienced forensic practitioners rated the extracts more consistently than did the inexperienced group. The level of variance across groups can be seen when comparing standard deviations for each group for each response; however, the source of the variance may be the extracts themselves and not just the listener. Even if groups rate consistently with respect to each other, the standard deviation may still be large if a high level variation is present within extracts (as demonstrated in §4.4.1). To avoid confusion between extract and participant variance, one mean score for each participant was calculated based on his/her scores for all extracts (illustrated in Figure 4-2 and Figure 4-3), and the standard deviation was produced by comparing participants' mean scores from the experienced group against those of the mean scores from the inexperienced group.

Table 4-3 reveals a lower mean distress categorisation score and a lower standard deviation value across both conditions for the experienced group. The experienced group tends to rate distress category lower than does the inexperienced group, i.e. the former group perceive less distress in the extracts, and their lower standard deviation indicates that the experienced practitioners tend to categorise material more consistently as a group than does the inexperienced group.

Table 4-3: Experienced and inexperienced group variance when categorising distress productions across both conditions

	Experienced group			Inexperienced group		
	with context	without context	average	with context	without context	average
Mean	2.14	2.34	2.24	2.39	2.55	2.47
S.D.	0.13	0.14	0.14	0.16	0.17	0.17

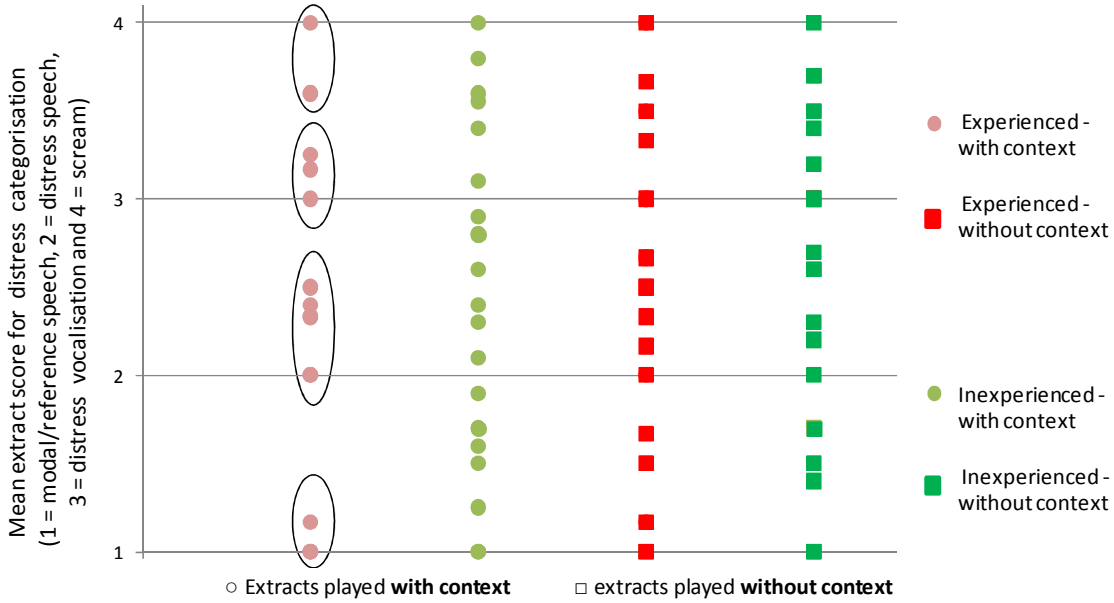
The same pattern is produced in the perceptible linguistic content ratings in Table 4-4. The experienced forensic practitioners exhibit a lower mean score and a lower standard deviation than their inexperienced counterparts (2.17 and 0.26 vs. 2.4 and 0.48, respectively). This implies that the experienced group assign more linguistic content to the material than do the inexperienced group, and that the experienced group are more consistent as a group when doing so.

Table 4-4: Experienced and inexperienced group variance when rating linguistic content across both conditions

	Experienced group			Inexperienced group		
	with context	without context	average	with context	without context	average
mean	2.00	2.33	2.17	2.19	2.61	2.4
S.D.	0.24	0.27	0.26	0.50	0.46	0.48

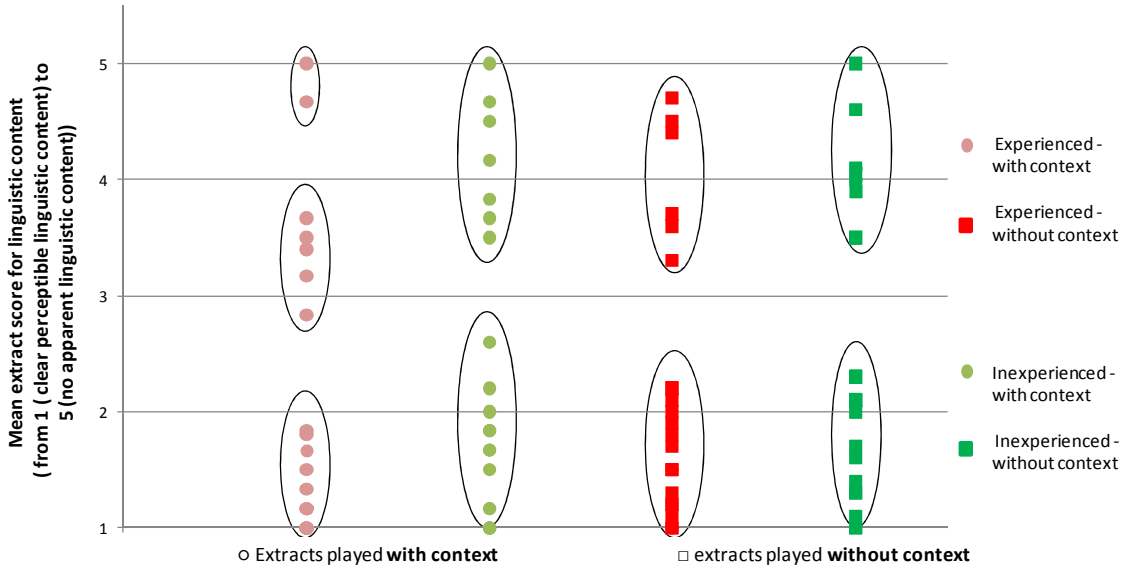
A further analysis of the data is presented in the scatter plots on the following page, which display the mean score per extract, grouped according to level of experience and condition.

Figure 4-14: Mean scores per extract by the experienced and inexperienced groups (red and green respectively) for distress categorisations across both 'with/without context' conditions. Each dot represents one extract.



When categorising distress, it can be seen that both groups use the full scale of the taxonomy in the 'without context' condition. This is repeated by the inexperienced group when extracts are heard with context, but the experienced group's mean scores are displayed as clusters, suggesting that the presence of context facilitates grouping comparable distressed productions together. Tight clusters are formed around distress categories 1, 3 and 4 (modal/reference speech, distress vocalisation, and scream, respectively) whereas a looser cluster is formed around distress category 2 (distress speech).

Figure 4-15: Mean scores per extract by the experienced and inexperienced groups (red and green respectively) for linguistic content ratings across both 'with/without context' conditions. Each dot represents one extract.



With respect to linguistic content ratings, in the 'without context' condition both experienced and inexperienced groups show a bipartite division, with extracts being perceived to either feature or lack linguistic content (represented by lower and higher scores, respectively). This pattern is also observed in ratings of extracts heard 'with context' by the inexperienced group, though some ratings approach the centre of the linguistic content scale, which suggests that the decisions are no longer as clear-cut as when extracts are heard without contextual information. The experienced group further demonstrate that in the 'with context' condition, the decision is no longer a binary one, since some extracts occupy a third, mid-way group, thus casting doubt on the level of perceptible linguistic content.

The 'with context' clusters displayed in both scatter plots support the hypothesis that the reduced amount of variation in participants' 'with context' scores is due to participants narrowing their choices of plausible interpretations for each extract, i.e. extract scores use the full range of the scale but are concentrated around certain points, forming clusters, as discussed in §4.4.2.

4.4.3.1 Summary of group performances

The experienced group assigned lower distress categorisation scores and higher linguistic content scores than did their inexperienced counterparts. It is hypothesised that experienced forensic practitioners perceive less distress and more linguistic content in the extracts than do those with less forensic experience. The experienced forensic practitioners also displayed smaller standard deviations in their scores, reflecting the fact that they rate more consistently as a group than do the inexperienced forensic phoneticians.

4.4.4 Participants' categorisation of data vs. original classification of data

To investigate whether the original taxonomy was easily and consistently applied across listeners using the same data in a replicable manner, it is important not only to examine the consistency of categorisations within and across participants, but also to look at the original classifications of the data before the experiment was conducted. (The original classifications refer to my classifications of the data made during my MSc research).

Table 4-5 shows the number and percentage of distress categorisation responses that match the original categorisation in the 'with context' condition, differentiated by group. Only responses pertaining to the 'with context' condition for distress categorisations are analysed here, since the original categorisation of forensic material by the author was based on a four-part scale and was classified after having heard the material several times with contextual information known.

Table 4-5: Amount of agreement between the original classifications by the author and the participants' categorisations of the same data.

	experienced			inexperienced			both groups		
Original categorisation	n	total	%	n	total	%	n	total	%
1 (modal/reference speech)	18	18	100	30	30	100	48	48	100
2 (distress speech)	26	59	44	50	98	51	76	157	48
3 (distress vocalisation)	17	48	35	43	89	48	60	137	44
4 (scream)	6	6	100	10	10	100	16	16	100

As previously mentioned in §4.4.1, there was no dispute concerning the categories from the peripheries of the taxonomy; discrepancies arose where extracts were categorised as either ‘distress speech’ or ‘distress vocalisation’ in the middle of the taxonomy. Extracts that were originally labelled ‘distress speech’ showed a greater degree of agreement (44% for the experienced group and 51% for the inexperienced group) than did the ‘distress vocalisation’ category (35% and 48% respectively). On the whole, the inexperienced group showed a greater degree of agreement with the original categorisation than did the experienced group (51% and 48% vs. 44% and 35%). This greater degree of agreement between myself and the inexperienced group is perhaps not surprising, since at the time of originally classifying the material, I would also have considered myself to be an ‘inexperienced’ forensic listener. Although I had had more exposure to distress material than the inexperienced listeners as a consequence of the theme of my MSc dissertation, I had no forensic casework experience at the time.

Table 4-6 provides further information detailing in the direction in which the discrepancies fell when listeners were about to categorising distress speech and distress vocalisations. The inexperienced group are, on the whole, just as likely to categorise the material as more distressing as they are to rate it as less distressing for both categories (distress speech was rated as being more distressing by 23% and less distressing by 26%, while distress vocalisations were rated as more distressing by 31% and less distressing by 28%). On the other hand, where the experienced group disagreed with the original classifications, they had a tendency to categorise the material as less distressing for both distress speech and distress vocalisation categories (distress speech was rated as more distressing by 14% and less distressing by 42%, and distress vocalisations were rated as being more distressing by 25% and less distressing by 52%).

Table 4-6: The direction of change where participants' categorisations differed from the original classifications by the author.

	experienced			inexperienced			both groups		
	n	total	%	N	total	%	n	total	%
participant score > original distress speech	8	59	14	23	98	23	31	157	20
participant score < original distress speech	25	59	42	25	98	26	50	157	32
participant score > original distress vocalisation	12	48	25	28	89	31	40	137	29
participant score < original distress vocalisation	25	48	52	25	89	28	50	137	36

A further analysis is presented in Table 4-7, which compares the participants' categorisations of distress against the original categorisations. The previous analysis compared the total number of extracts for each distress category with those from the participants as decided by myself, whereas the following analysis compares the total number of extracts for each distress category as classified by the participants and compared with my original classifications.

Extracts categorised as non-distress speech by participants were categorised originally in the same way in 49% of cases, with the other 51% being rated higher, i.e. as distress speech. For categories in the middle of the taxonomy, the original categorisations matched those of the participants most of the time (67% agreement for both distress speech and distress vocalisation). Where disagreements arose, the original classifications tended to be scored towards the middle of the taxonomy, i.e. being categorised higher for 'distress speech' but lower for 'distress vocalisation'. For the scream category, the original categorisations agreed with the participants' for just 28% of the responses, the remaining responses being categorised originally as less distressing, being labelled, for example, 'distress vocalisation'. The same trends were observed when comparing the scores from the experienced and inexperienced groups against my original classifications. However, the original responses appear more similar to those of the inexperienced group for categories in the least distressing part of the taxonomy (i.e. non-distress speech and distress speech), e.g. 55% agreement with the inexperienced group vs. 42 % agreement with the experienced group for non-distress speech, and 74% inexperienced group agreement

vs. 58% experienced group agreement for distress speech; and more like the experienced group for the more distressing part of the taxonomy (i.e. distress vocalisation and distress scream), e.g. 66% agreement with the inexperienced group vs. 71% agreement with the experienced group for distress vocalisations, and 26% inexperienced group agreement vs. 32% experienced group agreement for screams.

Table 4-7: Participants' distress categorisations versus original classifications by the author.

	experienced			inexperienced			both groups		
Categorisation	n	total	%	n	total	%	n	total	%
Original categorisation = participants' non-distress speech	18	43	42	30	55	55	48	98	49
Original categorisation > participants' non-distress speech	25	43	58	25	55	45	50	98	51
Original categorisation = participants' distress speech	26	45	58	50	68	74	76	113	67
Original categorisation < participants' distress speech	0	45	0	0	68	0	0	113	0
Original categorisation > participants' distress speech	19	45	42	18	68	26	37	113	33
Original categorisation = participants' distress vocalisation	17	24	71	43	65	66	60	89	67
Original categorisation < participants' distress vocalisation	7	24	29	22	65	34	29	89	33
Original categorisation > participants' distress vocalisation	0	24	0	0	65	0	0	89	0
Original categorisation = participants' distress scream	6	19	32	10	39	26	16	58	28
Original categorisation < participants' distress scream	13	19	68	29	39	74	42	58	72

These results suggest that the forensic data were originally categorised more cautiously, with many extracts occupying the middle of the taxonomy. The inexperienced group tend to follow this same pattern, whereas the experienced group are more confident in using the extremes of the taxonomy.

4.4.4.1 Summary of participants' categorisation of data vs. original classification of data

Disagreement occurred for categories in the middle of the taxonomy (i.e. distress speech and distress vocalisation) with the latter having the lowest degree of agreement between my classifications of the data (and therefore the expected classifications from the participants) and the participants' classifications. On the whole, the inexperienced group shared a greater degree of agreement between their classifications and mine than the experienced group. The experienced group are more likely to rate the extracts as containing less distress than did the original classifications, e.g. distress speech for distress vocalisation.

Although participants agreed with my original classifications of peripheral categories, I did not always agree with them when they classified extracts as belonging to these categories. Akin to the inexperienced phoneticians, my original classifications tended towards cautious categorising of the data using the middle of the taxonomy, whereas the experienced group were more willing to assign extreme categories to the data from the outset.

4.4.5 Discussion of results

In feedback following the experiment, it was reported that for certain extracts, the 4-part scale for distress categorisation was too restrictive, with some participants wanting to mark two categories. This, in fact, was the primary cause of responses being excluded from analysis. These extracts were perceived as straddling categories, reinforcing the criticism that the boundaries between categories are not clear-cut. The results confirmed that extracts representing the extremes of the taxonomy are easily categorised, but that those categories representing the middle of the taxonomy require greater delineation. This is further highlighted when viewing the scatter plot in Figure 4-14. Ideally, the dots representing mean extract values would cluster around the whole values of 1, 2, 3, and 4, signifying that each extract

was systematically rated as belonging to the same defined category by most speakers. This is not evident in the scatter plot, though the mean 'with context' scores for the experienced group do show signs of a clustering pattern.

A further limitation highlighted in participant feedback following the experiment was the apparent confusion between intelligibility and linguistic content. For the purposes of the experiment, 'linguistic content' referred to the perception of the productions as possibly speech-like, even if the listener could not discern what was said. In an effort to distinguish between intelligibility and linguistic content prior to conducting the experiment, the example of an unknown foreign language being unintelligible but carrying linguistic content (i.e. sounding speech-like) was provided. Despite this attempt to clarify the distinction between the two, for many participants one presupposes the other so that extracts perceived to be high in intelligibility on the categorisation scale (i.e. non-distress and distress speech) would be partnered with high linguistic content ratings (i.e. values 1 or 2 on the linguistic content scale). Participants were specifically asked to evaluate using these two scales in order to explore the relationship between the features of emotional arousal, intelligibility and linguistic content, which formed the basis of the original classification system. It was initially supposed that intelligibility and linguistic content would be easily differentiated, and it was hypothesised that if the broad mid-taxonomy categories with questionable intelligibility could be differentiated by presence or absence of linguistic content in the experiment, a modified taxonomy containing five categories in total (i.e. non-distress/reference speech, distress speech, distress vocalisation with linguistic content, distress vocalisation without linguistic content, distress scream) might prove to better describe the data. In hindsight, the distress categorisation scale could have employed different terminology in order to avoid preconceptions of our understanding of 'speech', 'vocalisation', and 'scream' clouding the categorisation classifications, and should have been redefined to focus on just one dimension. The limitations of the scales used in this experiment raise questions about which features ought to be considered when judging the level of distress. In some cases, distress can be universally recognised, as evidenced by the unanimous categorisation of extract 1 as a distress scream. In other cases, perceptible

distress may be reduced even when the source of the distress is made known, as in e.g. extracts 10 and 32 (§4.4.1).

4.4.6 Conclusions concerning the validation of the taxonomy

On the whole, the reliability and replicability of the original taxonomy demonstrated mixed results, and limitations of the taxonomy were discovered. Despite the anticipated variability of extracts and of individual performances, it was hypothesised that there would be a high level of cross-participant agreement between groups of similarly-trained forensic phoneticians and that ratings for the extracts would match my own original classifications. Although this was partially borne out - findings indicate that the experienced group tended to rate more consistently than the inexperienced group, and all ratings had above-chance levels of agreement with my own classifications - I consider the taxonomy in need of modification before it is implemented during the analysis of the data for the current investigation. The lack of agreement amongst participants in middle categories of the taxonomy and their lack of agreement with my original classifications of the peripheral categories demonstrate that the criteria used in the original taxonomy are insufficient for present purposes, and merit further development.

Consequently, a modified version has been put in place for the remainder of the analysis to be carried out for this investigation and classifications made previously in the authentic data are to be revisited. The major amendments to the taxonomy are as follows:

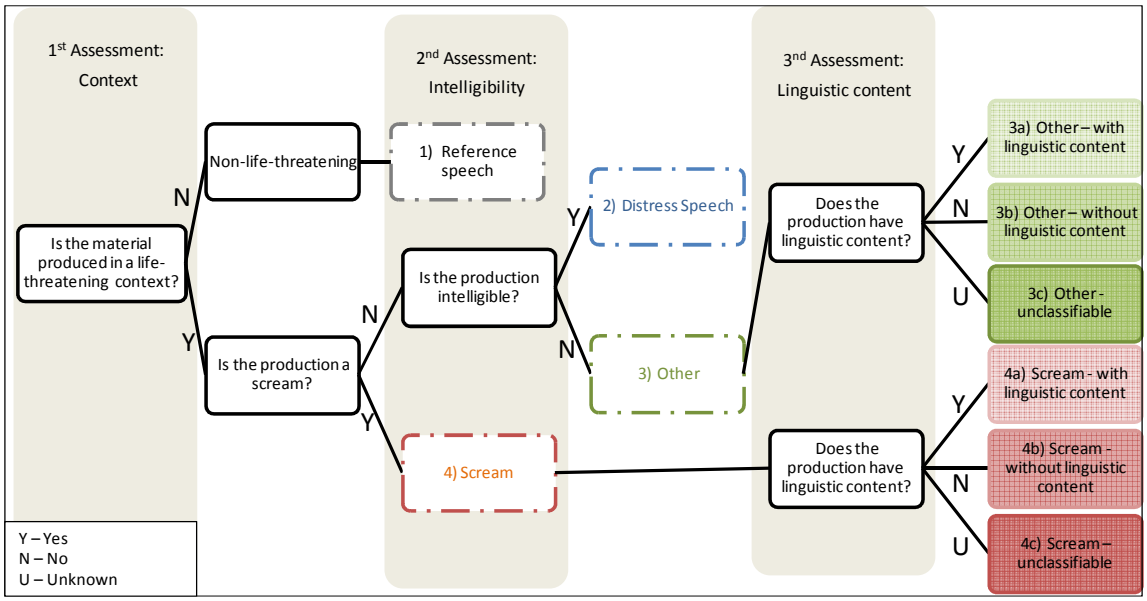
1. In recognition of (a) the fact that the use of ‘vocalisation’ for the category in between speech and scream had different interpretations by different participants and (b) no existing language terminology adequately describes a human vocal response which excludes speech and screams but includes “grey areas” such as productions where the presence of perceptible linguistic or paralinguistic features is questionable, the category is more aptly named ‘other’.
2. The notion of a continuum has been discarded, as it would be impossible to correctly infer speakers’ evaluations of what is more or less stressful in a distressing situation based on an audio recording. Moreover, if a continuum

model was employed the taxonomy would require a more consistent metric with which to frame definitions. Instead, only a categorical taxonomy of distress will be employed, with no reference to perceived levels of distress.

3. An improved metric is introduced which contains definitions based on three assessments. The first assessment is determined by the non-impressionistic criterion of whether the material was elicited in a distressing, life-threatening context. The second assessment is a (somewhat) subjective measure of whether the production was "intelligible". In the final assessment, a further subjective measure of presence or absence of linguistic content (or unknown) was added to aid distinctions within vocalisation and scream categories.
4. Revised definitions for the first three categories of the taxonomy now use a more consistent framework (see above point), though this will not be extended to the final category, 'scream', since no participant in the listening experiment rated screamed extracts as anything other than 'scream', despite the lack of definition.

The revised taxonomy is summarised as a decision tree in Figure 4-16 and it has already appeared in table form in the previous chapter (Table 3-4).

Figure 4-16: The revised taxonomy of distress.



4.5 Listening Experiment Results II - the influence of contextual information on listeners' perceptions of distress

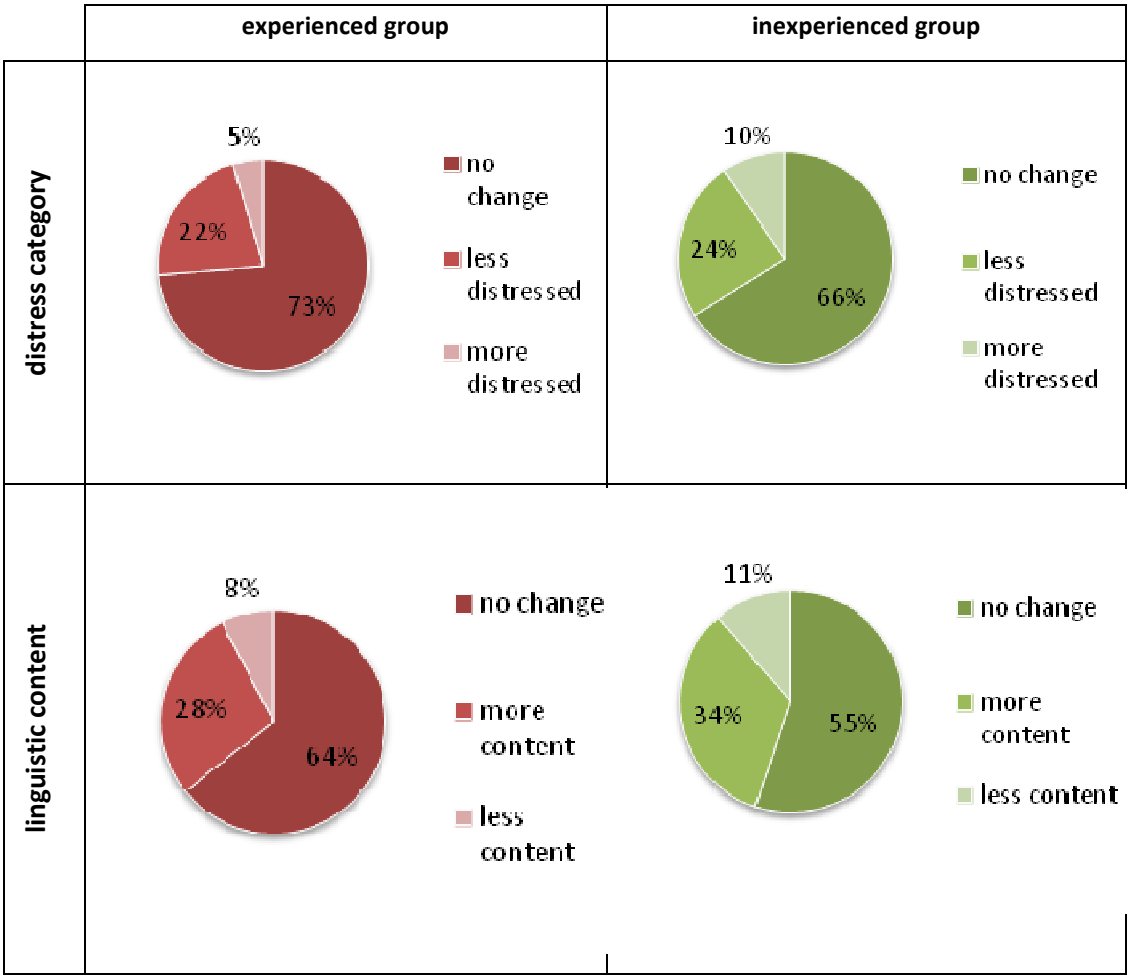
The following section investigates the influence of semantic context and case background on participants' ratings of forensic data. Although the previous section of this chapter demonstrates that the taxonomy used in the listening experiment is not without flaws, the responses from the listeners can still be used to indicate any changes (or lack thereof) in the listeners' judgments when hearing the extract in the presence and absence of context. The directionality of changes in their responses, if different from their original ones, will reveal whether contextual information affects listeners' perceptions of distress.

4.5.1 Changes in listeners' responses based on the presence or absence of contextual information

The pie charts below show the direction of the change from 'without context' responses to 'with context' responses for both distress categorisations and linguistic content ratings. Red represents the experienced group of forensic listeners, whereas green represents the inexperienced forensic listeners.

As illustrated, most responses remain the same, despite the change to the 'with context' condition; however, where participants altered their responses, extracts were categorised as reflecting lower degrees of distress and rated as having higher degrees of linguistic content when heard with contextual information. This pattern, as also demonstrated in §4.4.1 and §4.4.2, is evident for both groups of forensic listeners, though the experienced group were less prone to changing their responses when extracts were heard in the different conditions.

Figure 4-17: The change in direction from 'without context' responses to 'with context' responses for both the experienced and inexperienced forensic listeners (left and right columns, respectively) in terms of categorising the level of distress and linguistic content (top and bottom rows, respectively).



The pie charts show a simplified view of this trend since the denominator is based on the total number of responses for each scale and does not take into account the fact that some scores could not be marked higher or lower on the scale if they were already marked as end values. A more accurate representation of the trend is provided in Table 4-8 and Table 4-9 since the denominator has been recalculated to exclude scores that were already at the edge of each scale.[13] The pattern is clearly the

[13] This means that the total number of responses for which it is possible to rate an extract as belonging to a category of lower distress might be different from the total number of responses it is possible to rate as belonging to category of higher distress. For example, in Table 4-13 it can be observed that the experienced listeners rated 28 responses from a possible 92 as less distressed when hearing the extract with contextual information, resulting in 30% of the responses changing in this direction. For the same group of listeners, 6 responses were rated as being more distressed

same as that depicted in the pie charts, though the extent of the change in direction is more pronounced.

Table 4-8 Change in direction from 'without context' to 'with context' conditions in participants' responses categorising distress productions

Distress categorisation - both groups	**n**	**total**	**%**
with context score = without context score - *no change*	245	356	69
with context score < without context score (excluding 1) – *less distressed*	83	271	31
with context score > without context score (excluding 4) – *more distressed*	28	278	10
Distress categorisation - experienced	**n**	**total**	**%**
with context score = without context score - *no change*	95	129	74
with context score < without context score (excluding 1) – *less distressed*	28	92	30
with context score > without context score (excluding 4) – *more distressed*	6	107	6
Distress categorisation - inexperienced	**n**	**total**	**%**
with context score = without context score - *no change*	150	227	66
with context score < without context score (excluding 1) – *less distressed*	55	179	31
with context score > without context score (excluding 4) – *more distressed*	22	171	13

Table 4-9 Change in direction from 'without context' to 'with context' conditions in participants' responses rating linguistic content

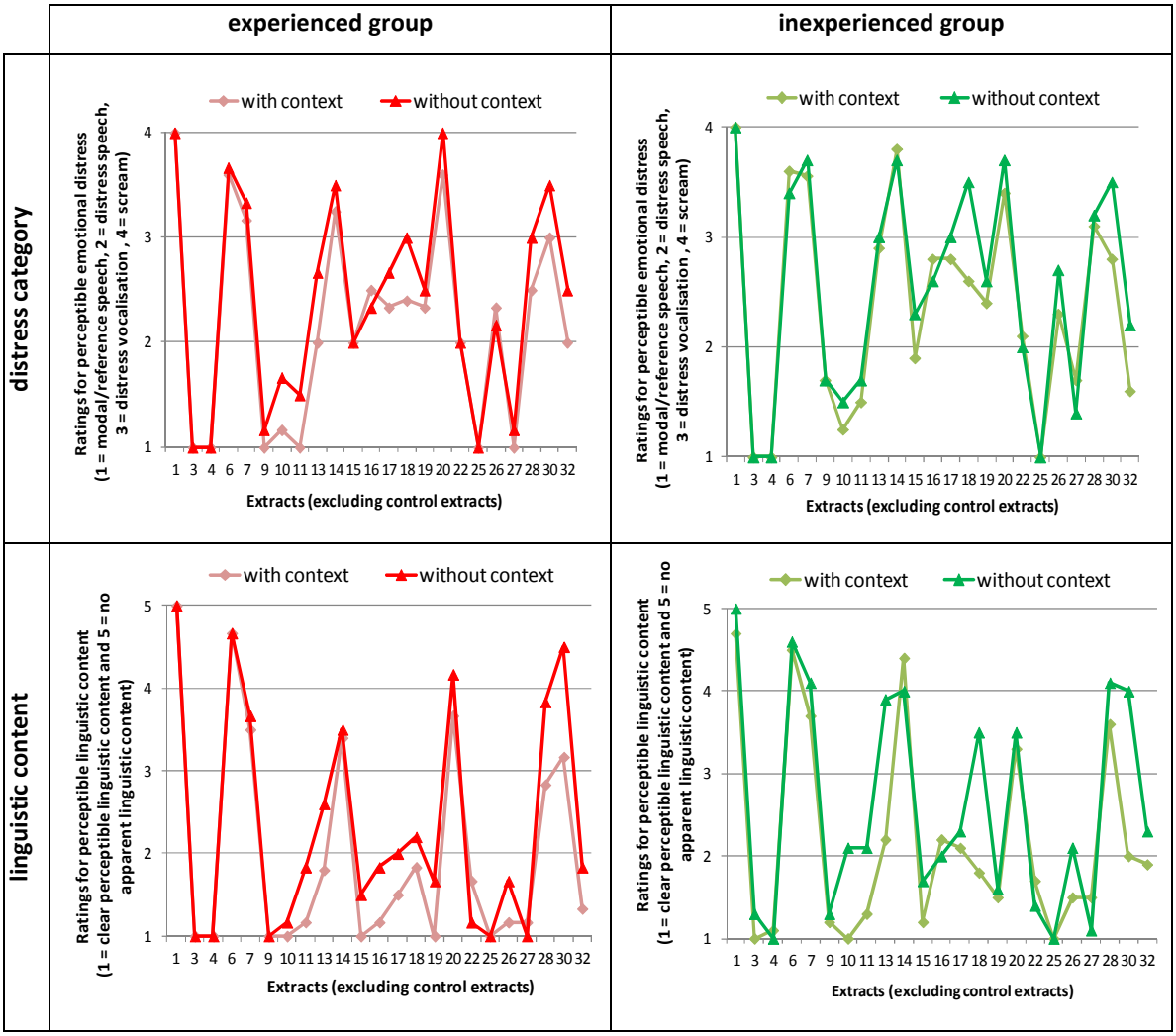
Linguistic content rating - both groups	**n**	**total**	**%**
with context score = without context score - *no change*	212	362	59
with context score < without context score (excluding 1) – *more content*	114	219	52
with context score > without context score (excluding 5) – *less content*	36	295	12
Linguistic content rating - experienced	**n**	**total**	**%**
with context score = without context score - *no change*	85	132	64
with context score < without context score (excluding 1) – *more content*	37	73	51
with context score > without context score (excluding 5) – *less content*	10	109	9
Linguistic content rating - inexperienced	**n**	**total**	**%**
with context score = without context score - *no change*	127	230	55
with context score < without context score (excluding 1) – *more content*	77	146	53
with context score > without context score (excluding 5) – *less content*	26	186	14

The effects of context and experience can also be illustrated in the following line graphs. These show the mean scores for distress categorisation and linguistic content

when heard with contextual information, out of a possible 107 responses. Therefore, 6% of responses resulted in a change of direction towards being more distressed.

rating per extract for both groups of listeners in both conditions. The same colour coding from the previous figure applies.

Figure 4-18: The average scores per extract for the experienced and inexperienced forensic listeners (left and right columns, respectively) for distress category (top row) and linguistic content (bottom row). Extract numbers are shown along the x axis and scores on the y axis.



Where only one line/dot is visible, it denotes that the scores overlap as there is no difference between the two conditions. For example, the first three extracts show that the distress category was rated the same when heard with and without context. Where a line is above or below the other, scores were rated as higher or lower on the scale, respectively. Lower scores on the distress categorisation scale represent less perceptible distress and lower scores on the linguistic content scale represent more linguistic content. Large spaces between each dot for the same extract represent a marked difference in rating between the two contexts.

On the whole, the line graphs show that, regardless of experience, both groups perceive lower degrees of distress and more linguistic content when extracts are heard with context (the pale lines tend to be below their darker counterparts), and the presence of context affects the experienced group of forensic practitioners in a more consistent way than it does the inexperienced group, since there are fewer crossovers in the lines for the former group.

ANOVA showed a significant effect with respect to both the distress production categorisations and the linguistic content ratings for context and experience, with context exercising the stronger effect. The distress productions categorisations showed for context: $F(1,707) = 6.29$, $p < .012$; and experience: $F(1,707) = 4.58$, $p < .033$. The linguistic content ratings showed for context: $F(1,714) = 11.94$, $p < .001$; for experience: $F(1,714) = 5.24$, $p < .022$). The interaction between experience and context was not significant for either of the two scales. Due to the small and unbalanced number of participants in each group, age, sex and native language were not included in the statistical analysis. The effect of order, i.e. whether the extract was first heard with or without context, was not significant.

4.5.2 Discussion of results

The second aim of the experiment was to see whether exposure to semantic context and case background affect the listener's perception of the extract and consequently lead to a change in categorisation. A significant effect of context was predicted and found. The presence of context was expected to help resolve issues concerning degree of perceptible linguistic content in the extracts since top-down information allows the listener to narrow down the number of possible and likely interpretations of a production, though this approach may have consequences for transcriptions involved in criminal investigations.

Surprisingly, the presence of context lowers the distress categorisation score so that extracts heard without context are perceived as containing more distress. It was anticipated that once the participant knew more of the (often) horrific circumstances surrounding the victim's attack, the production would be perceived as having higher degrees of distress. However, the participants were primed to hear dreadful

productions from victims since the violent nature of the recordings was not concealed from them and was, indeed, referred to in the instructions documentation should any participant subsequently choose not to proceed. One hypothesis is that the extracts played 'without context' represent the unknown and so what is unknown may not only be perceived as being more distressing, but also more distressed. Furthermore, the priming of violent audio extracts may have fired up the participants' imagination to the extent where the true circumstances revealed in the 'with context' extracts were not as horrific as first anticipated. On the other hand, the context can hinder categorisation of distress, particularly when the production is delivered without the vocal cues to distress we may expect (e.g. high F0). This mismatch results in an increased variation in scores for the production and demonstrates further the inter-victim variation in speech when responding to situations of distress.

A further consideration is that the experienced group have heard many forensic recordings and are more familiar with the associated problems such as the variability of quality, duration and amount of available speech. All have heard violent attack material previously, though some are more exposed to material of this nature than others. Given their familiarity with the material, they may apply the scales for distress and linguistic context differently as they have more reference material on which to base decisions. Furthermore, they are, on the whole, older than the inexperienced group and so their longer life experience may also provide increased opportunity for exposure to distress material in their day-to-day lives. The inexperienced group, on the other hand, had had either limited or no exposure to victims' productions during violent attacks and so their knowledge of distress material might be derived from how they expected people in distress should sound, perhaps based on portrayals in TV and film as well as their growing reference sample from increased exposure to the experiment. The amount of intra- and inter-victim variation was perhaps not expected and the inexperienced group may have experienced more of a "shock factor", thus perceiving the extracts as containing more distress and less linguistic content. In addition, their lack of experience might predict a lesser degree of 'calibration' than the experienced group who work closely together and have analysed a wide range of cases. A further consideration is that the

experienced group may be more inflexible in their decisions due to their increased awareness of and familiarity with forensic material. As a result, the experienced group have a higher degree of agreement amongst themselves when assessing extracts.

Since the stimulus extracts all originated from an archive at JPFA, some of the experienced group are likely to have heard parts of the material before. Indeed, most of the experienced group identified at least one extract from a well-known case amongst the data. Where relevant, some members of the experienced group noted which extracts they thought they recognised so that these responses could be analysed further at a later date if required. The design of the experiment (multiple blocks, randomly ordered, 2 conditions) should have minimised potential memory effects, though it has to be recognised that some of the experienced group members may have rated and categorised extracts in the 'without context' condition while having knowledge about the case, especially if the 'with context' condition was presented first, providing details which match and confirm their memory of the case details.

The significant effect of context on perceptible linguistic content lends itself to the debate concerning how best to approach transcribing forensic material for legal purposes. On the one hand, transcriptions ought to be as objective as possible, especially when used for judicial purposes (Fraser 2003). To avoid preconceptions entering into the transcription, a 'bottom up' approach, i.e. one whereby the listener first undertakes the task without contextual information, is encouraged (Fraser 2003). However, findings from this experiment highlight the fact that the recovery, and hence attribution, of linguistic content on the basis of a victim's brief production is unlikely to be achieved by considering solely the internal properties of the sound. Higher-order information – including semantic context and background story – must play a pivotal part in the interpretation.

4.5.3 Summary of influence of contextual information on listeners' perceptions of distress

The presence of listener experience as well as context prove to have a significant effect on participants' responses when both categorising distress productions and rating linguistic content. Extracts heard with context were categorised as having less perceptible distress and more perceptible linguistic content. These findings are of particular relevance to those debating the best approach to transcription of forensic material.

4.6 Chapter summary

This chapter recounted a perceptual experiment that aimed to test the reliability and replicability of the distress taxonomy(Roberts 2008) as well as to investigate the influence of contextual information on listeners' perceptions of distress. Findings from this experiment reveal that:

- the taxonomy was applied reliably and replicated with some success, though modifications to the taxonomy were proposed and implemented before further analysis took place. Consequently, findings from the analyses reported in the next chapter employ the modified taxonomy (described in §4.4.6 and §3.3.3).
- contextual information and forensic experience proved to have significant effects on listeners' perceptions of distress. Extracts heard 'with context' were categorised as exhibiting less perceived distress and more linguistic content than those that were heard 'without contact'. Experienced forensic listeners were more consistent in their judgments than were inexperienced forensic listeners.

5. Findings of Acoustic Study

This chapter reports the results of a study of the properties of the authentic and acted distress recordings using the acoustic analysis techniques presented in chapter 3. It contains five sections. The first four sections each concern an acoustic parameter (fundamental frequency, intensity, tempo and vowel formants, respectively). The final section summarises the acoustic findings. Each parameter is described in general, on an aggregate level, and then represented according to vocal response categories as defined by the taxonomy proposed in §3.3.3. All parameters are discussed in terms of the three principal lines of comparison: distress material versus reference material from victims; distress material versus reference material from actors; and distress material from actors versus distress material from victims. For comparisons of distress between actors and victims, the actors' distress refers to their rehearsed distress material.

5.1 Fundamental frequency

Results for F0 are illustrated in the following subsections as bar charts, boxplots and scatterplots using logarithmic Hertz (Hz) scales and semitones (ST). Although a linear display of F0, as shown in Figure 1, helps visualise the changes in the physical rates of vocal fold vibration (e.g. if a hypothetical speaker has a mean fundamental frequency of 100Hz in reference speech but 400Hz when in distress, we can deduce that the vocal folds are vibrating four times as fast), the perceived pitch of periodic sound is not linear (Nolan 2003). A fast vocal fold vibration will result in a high F0 and will have a high pitch (the perceptual correlate of F0). Similarly a low rate of vocal fold vibration will have a low F0 and a low pitch. However, despite this relationship, the perceptual difference between higher frequencies is not equivalent to that between lower frequencies. From psychophysical studies, it is known that pitch is perceived in an approximately linear fashion for F0 values below 500 Hz, but in a logarithmic fashion for F0s above that level (Jongman et al. 2006: 209). The presentation of F0 results using a purely linear scale risks overemphasizing the higher frequencies, whereas a logarithmic scale may overemphasize the lower frequencies (Fant 1971: 241). F0 comparisons amongst actors and victims contain material ranging from 74 Hz (Actor 2's minimum F0 in reference speech material) to

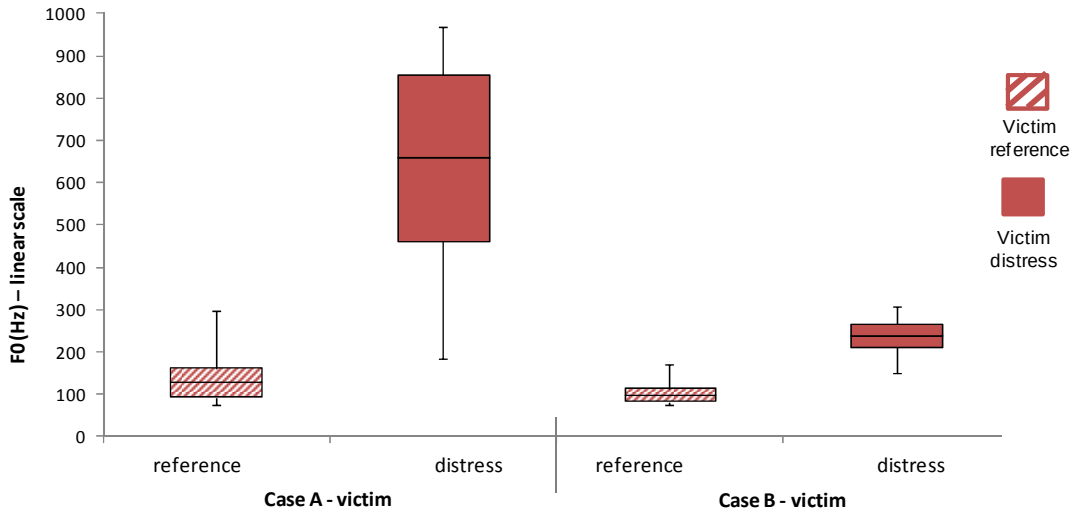
1580 Hz (Victim D's maximum F0 when screaming), and so results are illustrated using logarithmic scales. In addition to logarithmic values in Hertz, some F0 comparisons are shown in perceptual units of semitones (ST) - small, logarithmic intervals between two frequencies (12 semitones form an octave) - as a useful relative approach to comparing across male and females. Both a semitone and logarithmic Hz scale are used since, although both represent changes in vocal fold vibration, the conversion from Hz to ST does not best show the (in some cases) dramatic increases in F0 in the same way that Hz can.

5.1.1 Reference vs. distress in victims

As noted in §3.2.1, only two cases contained both reference and distress material: Cases A and B. In both cases the victims were male and had received a serious injury in a violent attack and had then received fatal second wounds. The distress speech material from both victims is taken from the period between their first and final injuries.

Displayed below are adapted boxplots showing absolute F0 mean (horizontal bar), standard deviation (box surrounding horizontal mean bar), minimum and maximum (lower and upper points of the whiskers, respectively). As illustrated in Figure 5-1, the change in F0 from reference to distress material is greater for Victim A (mean F0 of 659 Hz in distress) than for Victim B (mean F0 in distress is 239 Hz). Both men's reference material is within the typical F0 range of adult males, taken to be around 120 Hz (Fry 1979: 68).

Figure 5-1: Adapted boxplots showing absolute F0 mean, S.D., min. and max. for male victims in reference (hatched red) and distress (block red) conditions using a linear scale.



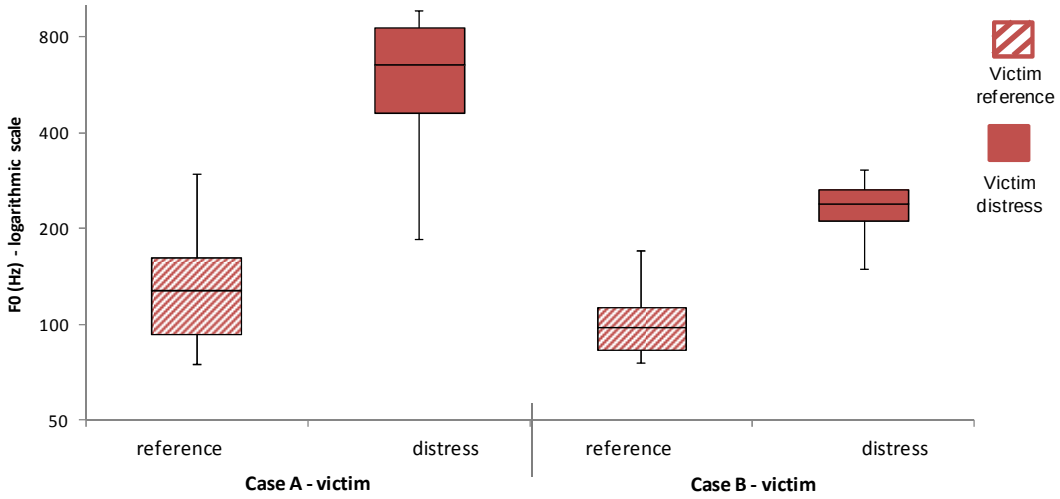
From Figure 5-1, it appears that F0 variability (represented as standard deviation - S.D.) increases in both victims' distress conditions, especially for Victim A. However, with an increase in F0 it is mathematically to be expected that absolute standard deviation will also increase (Jessen 2009: 129). If we therefore express standard deviation in semitones (Table 5-1), we see that Victim A has a similar range of pitch variation in both reference and distress conditions with distress production being slightly more variable (9.6 ST vs. 10.7 ST, respectively) whereas Victim B shows a lower range of pitch variation in distress (4.0 ST vs. 5.3 ST).

Table 5-1: F0 mean, min., max., and S.D. for victims in reference and distress conditions.

Victim	Condition	Mean (Hz)	S.D. (Hz)	Min. (Hz)	Max. (Hz)	S.D. (ST)
A	reference	127	34	75	299	9.60
	distress	659	198	185	968	10.71
B	reference	98	15	76	171	5.32
	distress	239	27	150	307	3.95

Both victims show within-speaker consistency in pitch variability, i.e. the victim who is more variable in reference speech is more variable in distress speech. This is well-illustrated in Figure 5-2, which reproduces the boxplots on a logarithmic scale.

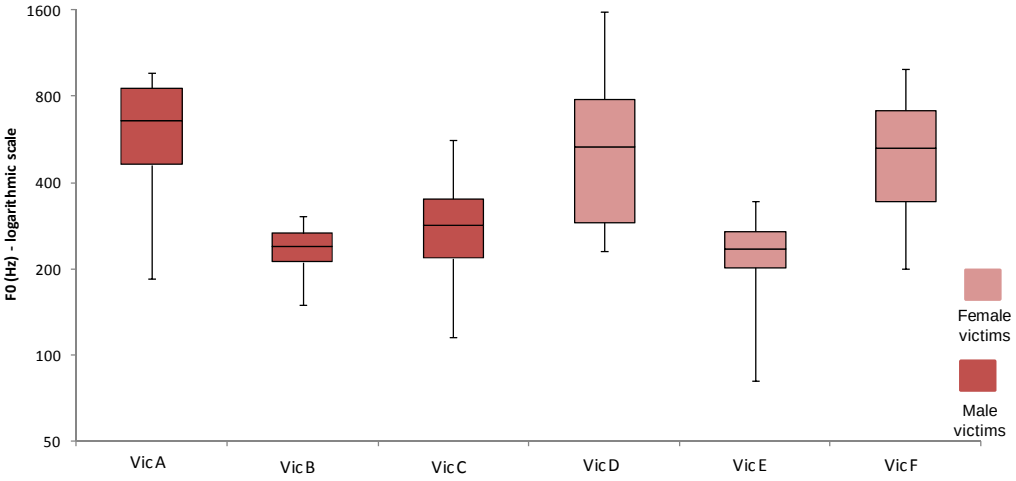
Figure 5-2: Adapted boxplots showing F0 mean, S.D., min. and max. for male victims in reference (hatched red) and distress (block red) conditions using a logarithmic scale.



As illustrated by the previous table and figures, Victim A produces a more dramatic increase from reference to distress material, increasing 28.5 semitones from one condition to another, whereas victim B increases by just 15.2 semitones.

Although there is no reference material available for Victims D-F, boxplots showing their absolute F0 means, standard deviations, minima and maxima in distress speech are displayed below for comparison with Victims A and B.

Figure 5-3: Adapted boxplots showing F0 mean, S.D., min. and max. for all victims (males in dark red, females in light red) in distress conditions using a logarithmic scale.



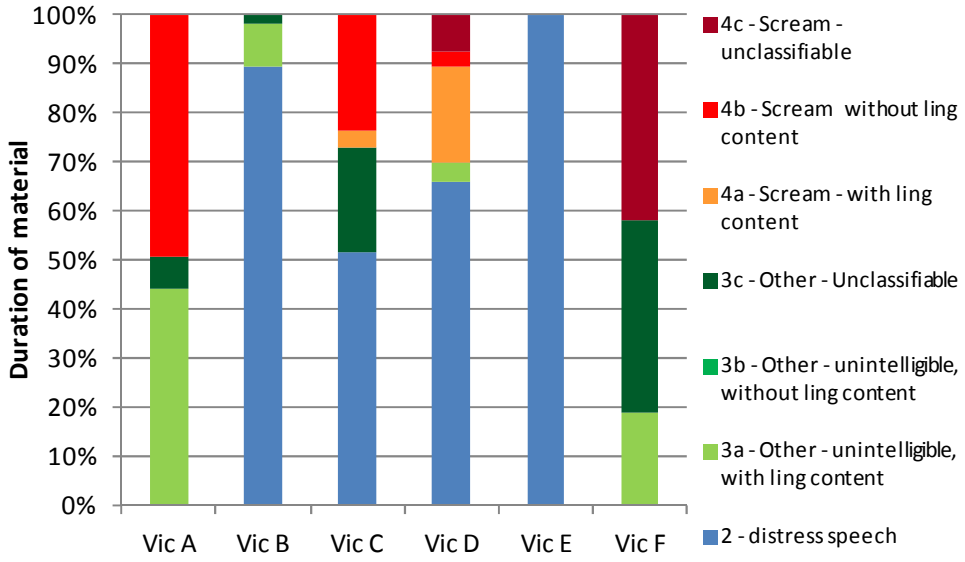
Interestingly, there is no clear division in Figure 5-3 and Table 5-2 between the male and female victims; male Victim A has a mean and maximum F0 similar to those of female victims D and F, whereas female Victim E falls within the range of the male victims B and C.

Table 5-2: F0 mean, S.D. (Hz), min, max. and S.D. (ST) for all victims in distress speech.

Speaker	Mean	S.D. (Hz)	min	max	S.D. (ST)
VIC A	659	198	185	968	10.7
VIC B	239	27	150	307	4.0
VIC C	284	67	116	563	8.3
VIC D	537	246	230	1580	17.1
VIC E	236	33	82	343	4.9
VIC F	529	18	201	994	12.7

Turning our attention to the manner of the victims' vocal responses, Figure 5-4 illustrates the proportion of the victims' speech as classified by the taxonomy in §3.3.3. Victims A and F produce a combination of screams (with unclassifiable or no perceptible linguistic content) and other vocalisations (some unclassifiable, and some unintelligible but with suspected linguistic content), whereas Victims B and E produce mainly intelligible distress speech, with a short unintelligible production that was unclassifiable in the case of Victim B. The other two victims (C and D) show almost the full spectrum of distress categories, with speech, screams and 'other' all represented.

Figure 5-4: Proportion of different manners of vocal response from the total distress material for all victims using the distress taxonomy.



The mean F0 of each category varies across all speakers, as shown in Figure 5-5. On the whole, all screamed categories tend to have a higher F0 than 'other' (vocalisation) categories, which in turn have a higher F0 than distress speech. The increase across the spectrum of categories is particularly well highlighted by Victims C and D. The pitch variability, i.e. standard deviation of each category (Figure 5-6), appears to follow a similar, yet not fully consistent, pattern of a gradual increase from distress speech to distress 'other', and then finally scream. All things being equal, we might expect the victims' reference speech to have a smaller standard deviation than their corresponding distress ones. However, this will depend on style of reference speech (e.g. monotone or not) and scream (e.g. level pitch or contour). Without more reference material available, it is hard to generalise differences between reference and distress standard deviation of F0.

Figure 5-5: Mean F0 for categorised vocal responses for each category of the distress taxonomy across all victims using a logarithmic scale.

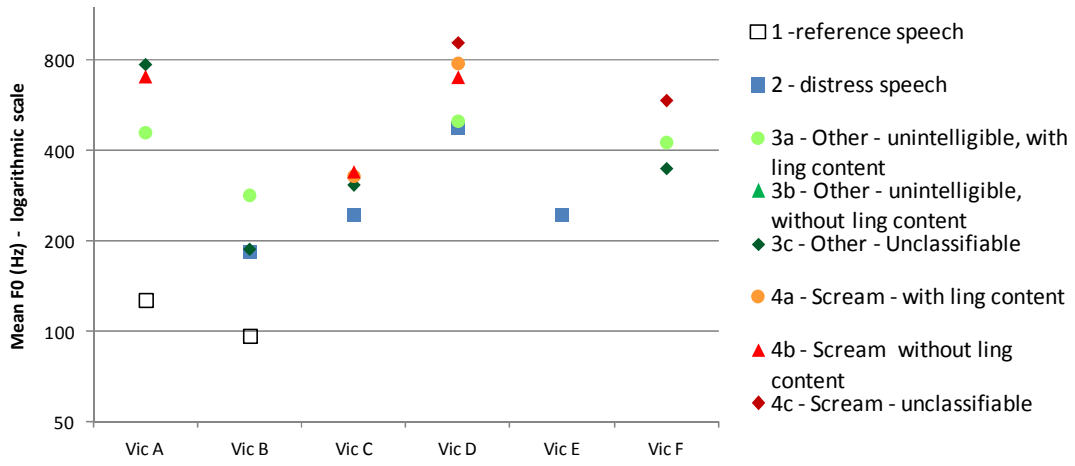
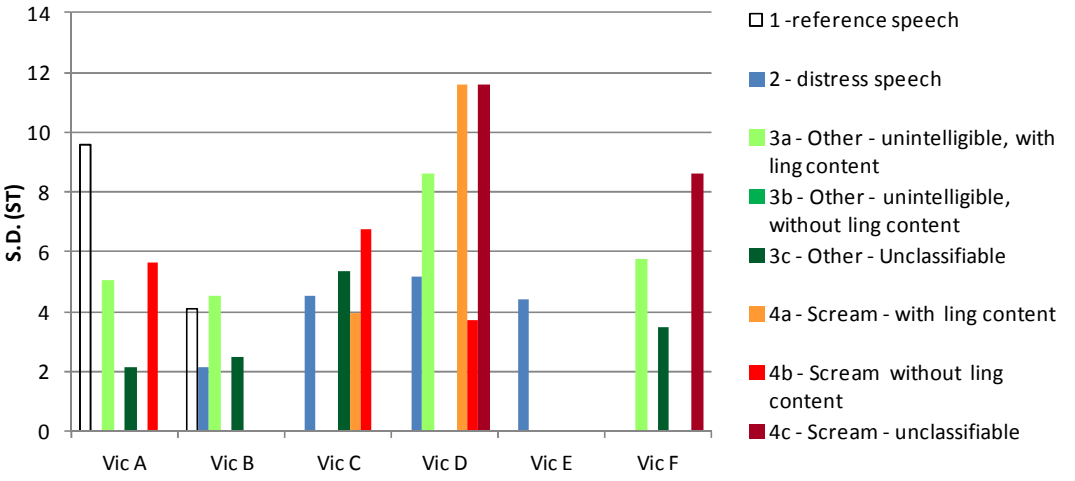


Figure 5-6: Standard deviation of F0 for categorised vocal responses for each category of the distress taxonomy across all victims using a logarithmic scale.



Since reference material is available for Victims A and B, the next paragraph will focus further on their F0 data, which are summarised in Table 5-3. As expected from the data represented in Figure 5-2 and Figure 5-3, Victim A's mean F0 shows a greater increase than that of Victim B, with unintelligible vocalisations containing perceptible linguistic content produced at around 460 Hz, and screams and unclassifiable vocalisations produced between 700 and 800 Hz. Victim B's distress speech and unclassifiable vocalisation both have an F0 of around 185 Hz. For victim A, the pitch variability within each category is fairly consistent, with standard deviations of approximately 5 ST in reference speech, screams, and unintelligible

productions with linguistic content, though unclassifiable productions varied only within 2 ST. Victim B's pitch variability decreased in distress conditions with a range of approximately 2 ST versus 4 ST in reference material.[14]

Table 5-3: F0 mean and S.D. across distress categories by victims A and B.

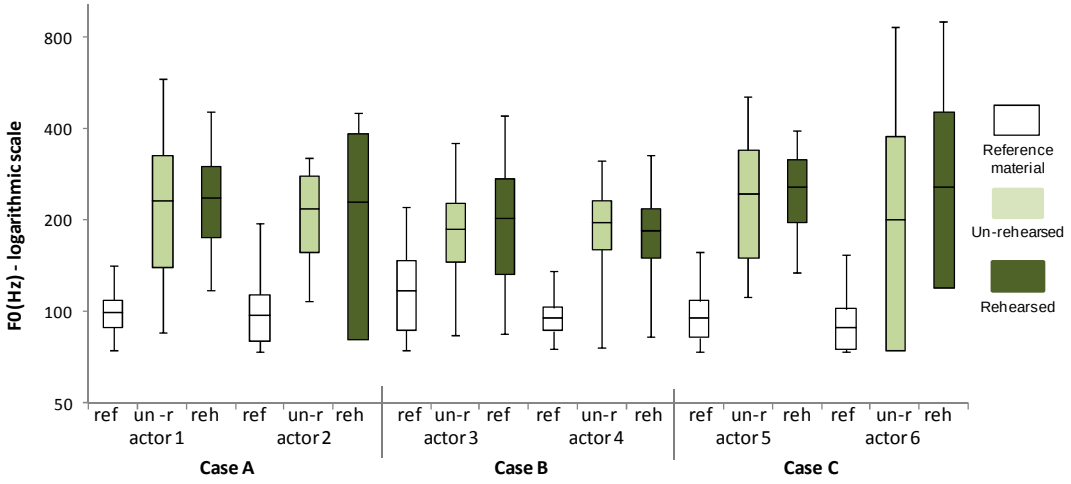
	Victim A		Victim B	
	F0 (Hz)	S.D. (ST)	F0 (Hz)	S.D. (ST)
1 - reference speech	127	5.04	97	4.09
2 - distress speech	x	x	183	2.11
3a - Other - unintelligible, with ling content	459	5.04	x	x
3b - Other - unintelligible, without ling content	x	x	x	x
3c - Other - Unclassifiable	775	2.14	187	1.94
4a - Scream - with ling content	x	x	x	x
4b - Scream without ling content	706	5.63	x	x
4c - Scream - unclassifiable	x	x	x	x

5.1.2 Reference vs. distress in actors

F0 data for actors is presented to compare their reference material with both their rehearsed and unrehearsed distress material. There is an increase in F0 mean and standard deviation for all actors from reference speech to distress speech conditions. Figure 5-7 shows the change in mean F0 and standard deviation (Hz) in male actors across all three conditions. On the whole, the increase in mean F0 in rehearsed and unrehearsed conditions are strikingly consistent, but the increase in standard deviations varies.

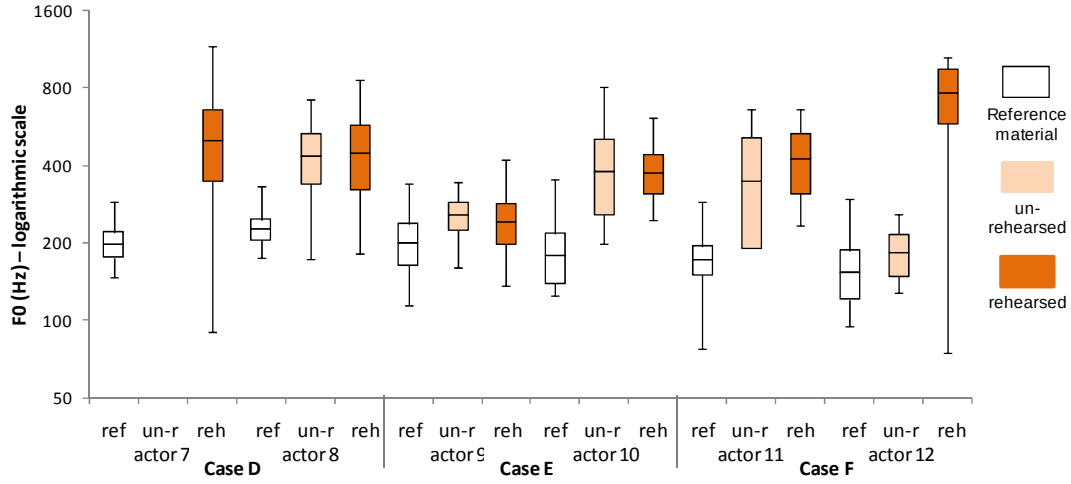
[14] It should be noted that the mean F0 and standard deviation per category differ from values reported when comparing across conditions on an aggregate level. This is due to the equal weighting of averaged, categorised vocal responses - they are not normalised to take into account the proportion of each category from the total duration of the material.

Figure 5-7: F0 mean, S.D., min. and max. for all male actors in reference and distress conditions (reference in white, unrehearsed in pale green, rehearsed in dark green) using a logarithmic scale.



A similar picture emerges with the female actors. Figure 5-8 shows that F0 means and standard deviations increase from reference speech material to rehearsed and unrehearsed distress conditions, though the extent of the increase varies across actors. Actor 12, however, does display a noticeable difference between the increase in the rehearsed and unrehearsed conditions.

Figure 5-8: F0 mean, S.D., min. and max. for all female actors in reference and distress conditions (reference in white, unrehearsed in pale orange, rehearsed in dark orange) using a logarithmic scale.



For all actors, the F0 varied significantly across the three conditions (χ^2 (2) = 17.63, p < 0.001). Wilcoxon tests were used to follow up this finding and a Bonferroni

correction applied so that all results are reported at a 0.0167 level of significance (0.05/3). F0 increased significantly in rehearsed distress performances compared with reference passage material (Wilcoxon $Z = 3.059$, one-tailed, $p < 0.001$, $r = 0.62$). Likewise, F0 values were significantly higher in unrehearsed distress performances for all actors than in control reading passage material (Wilcoxon $Z = 2.934$, one-tailed, $p < 0.0015$, $r = 0.63$). There was no statistically significant difference between the actors' rehearsed and unrehearsed mean F0 values.

With the exception of Actors 4 and 9, mean F0 was higher in the rehearsed distress condition than the unrehearsed condition, though it should be noted in most cases the mean F0 values for both distress conditions are very similar (only Actors 6 and 12 demonstrate a noticeable contrast between the two conditions). The increase in semitones from reference to distressed conditions is shown in Figure 5-9 and Figure 5-10, and it illustrates well the similar increase in level of mean F0 in both conditions. Similarly, there was no significant difference between the actors' rehearsed and unrehearsed increase in F0 (in semitones), nor between males' and females' increase in F0. From descriptive statistics, the females are more variable in their increase, though this may be due to their greater F0 range.

Figure 5-9: Increase in semitones from reference to distress conditions for male actors.

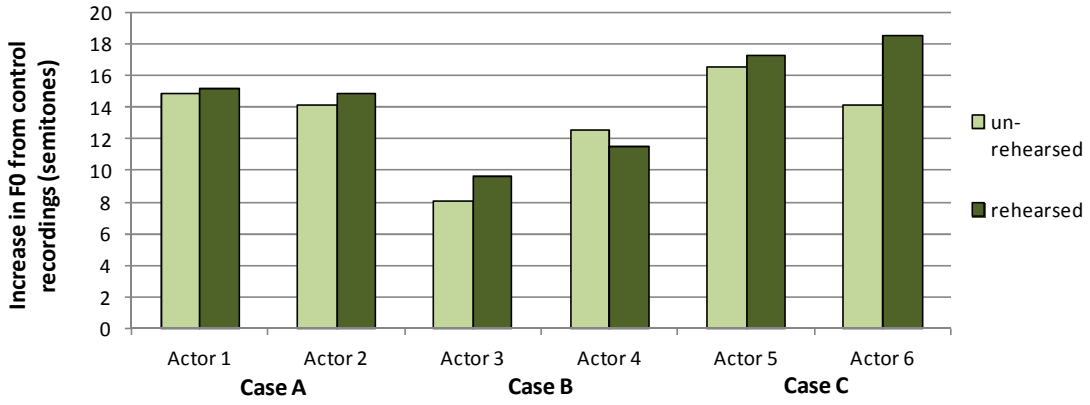
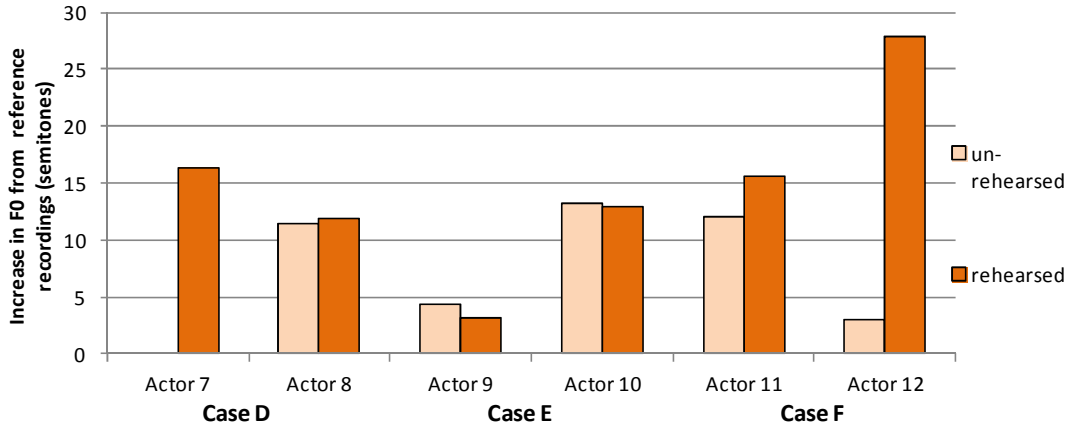


Figure 5-10: Increase in semitones from reference to distress conditions for female actors.



Unsurprisingly, the pitch variability as measured by standard deviation in semitones also increases across all actors between un-rehearsed and rehearsed conditions, although differences are apparent between the conditions (Appendix F1). Three male actors (1, 5 and 6) show greater variability in the unrehearsed condition, whereas Actors 2 and 3 are more variable in the rehearsed condition. Actor 4 is level for both distress conditions. Among the female actors, 8, 9, and 12 show a greater variability in the rehearsed condition compared with actors 10 and 11, who display more variability in the unrehearsed condition.

For all actors, the standard deviations did not significantly change across the three conditions (χ^2 (2) = 3.49, $p > 0.05$). However, Wilcoxon matched pairs tests were conducted with a Bonferroni correction applied (all effects are reported at 0.0167 level of significance) and they showed that the standard deviation significantly increased in the rehearsed distress condition when compared to corresponding reference passage material (Wilcoxon Z = 2.432, two tailed, p <0.015, r = 0.5). Similarly, S.D. significantly increased in unrehearsed distress performances for all actors compared with control reading-passage material (Wilcoxon Z = 2.401, two tailed, $p < 0.016$, $r = 0.51$). There was no statistical difference between the rehearsed and unrehearsed standard deviations.

The manner of the actors' distress responses using the taxonomy are illustrated in Figure 5-11 and Figure 5-12. On the whole, there was an upshift in the distress categorisations in actors' rehearsed condition. For all actors, the majority of responses were categorised as distress speech and therefore as intelligible, unlike

some of those for victims. Among the male actors, Actors 1 and 2 also make a lot of use of the 'other' category, especially vocalisations which are perceptibly unintelligible and without linguistic content. Screams are produced by four male (Actors 2, 4, 5 and 6) and all are without perceptible linguistic content. Female actors also have a higher proportion of intelligible distress speech than the other distress categories and appear to make more use of the 'other' category in the taxonomy. In contrast with the male actors, the females do make use of other vocalisations with linguistic content as well as without. Screams do not feature as much as for the male actors. Only Actors 7 and 8 produced screams, and when they did, they were screams containing perceptible linguistic information. However, there was no significant change in the proportion of use of specific categories among the actors.

Figure 5-11: Proportion of different manners of vocal response from the total distress material for all male actors.

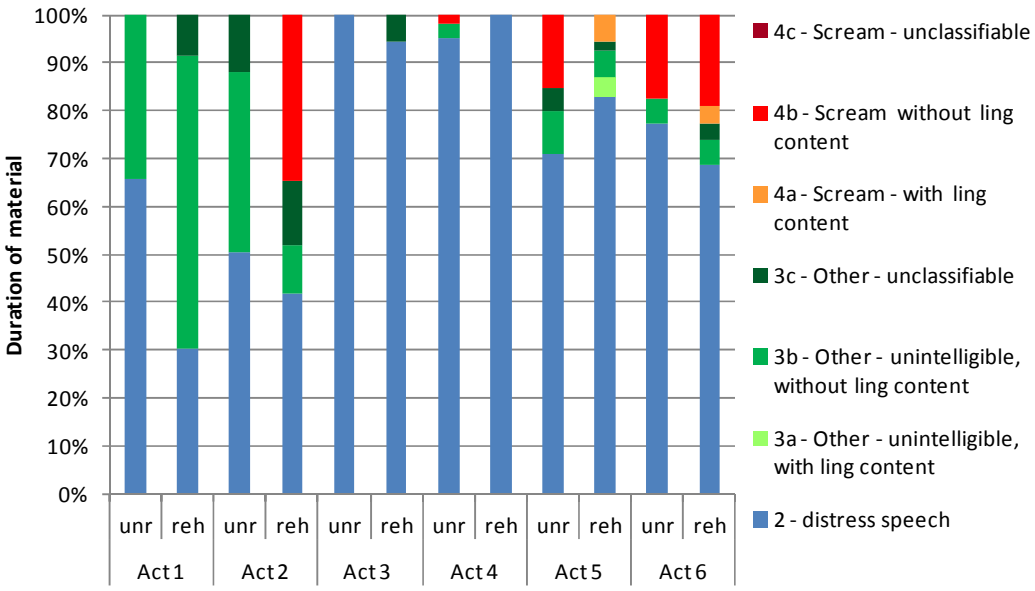
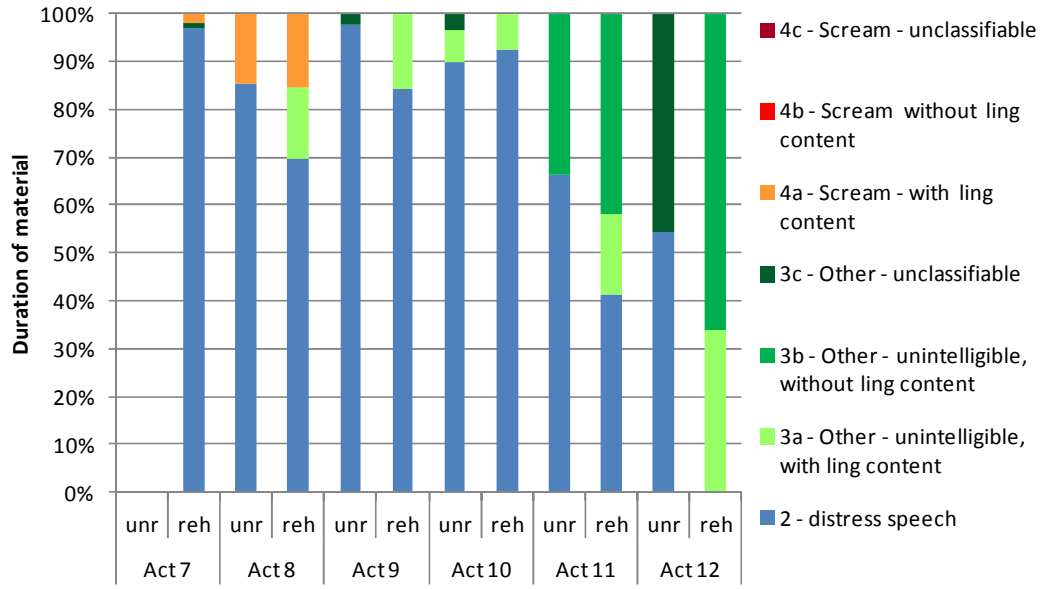


Figure 5-12: Proportion of different manners of vocal response from the total distress material for all female actors.



The breakdown of mean F0 for each actor in rehearsed and unrehearsed distress conditions for all the categories of the taxonomy is displayed in Figure 5-13 and Figure 5-14. All distress categories (2, 3a-c, 4a-c) show visible increases in mean F0 relative to the reference material. Only Actor 2's and Actor 9's mean F0 values for distress speech and unclassifiable vocalisations are slightly lower than their reference speech. For Actor 2, this is possibly due to his performing part of the script in a breathy whisper (Laver 1980), feigning drifting into unconsciousness. A tentative pattern can be seen in that, in most actors (Actor 1 is a noticeable exception), the increase in mean F0 is greater when screamed, and lower in distress speech. Vocalisations classified as 'other' tend to be intermediate in terms of F0 increase.

Figure 5-13: Mean F0 for categorised vocal responses for male actors.

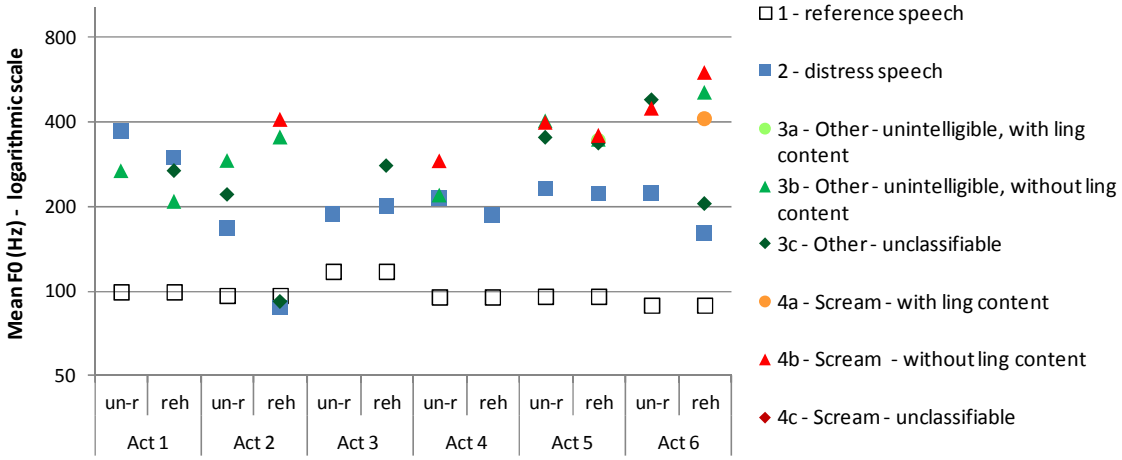


Figure 5-14: Mean F0 for categorised vocal responses for female actors.

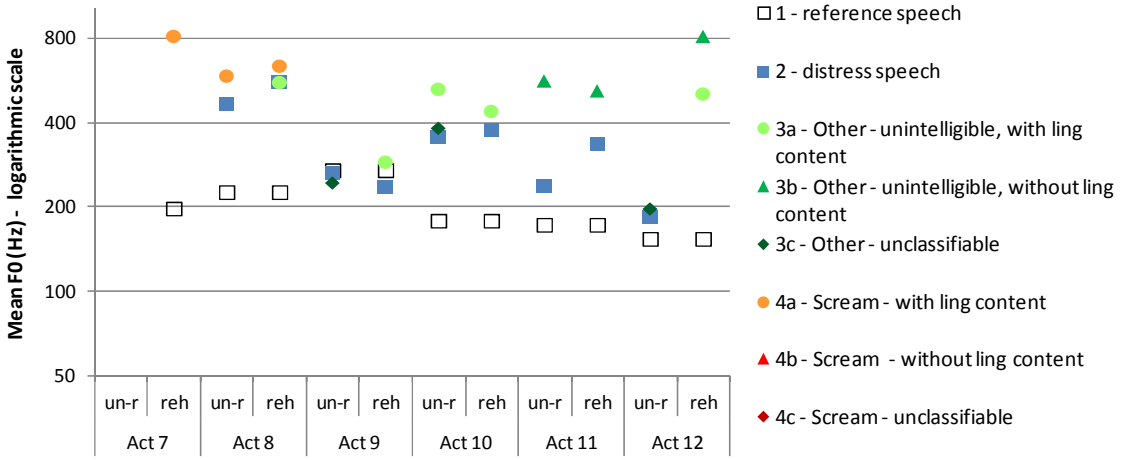


Figure 5-16 shows pitch variability for each category across speakers. On the whole, for male actors, those who show more F0 variability in reference speech are also more variable in distress speech. Furthermore, distress speech tends to be the most variable distress category. In all but Actor 6, 10 and 11's productions, other vocalisations and screams are less variable than distress speech, and in some cases even less so than their reference speech, e.g. for Actors 2, 3, 5 and 12.

Figure 5-15: Standard deviation of F0 (semitones) for each category of the distress taxonomy in male actors.

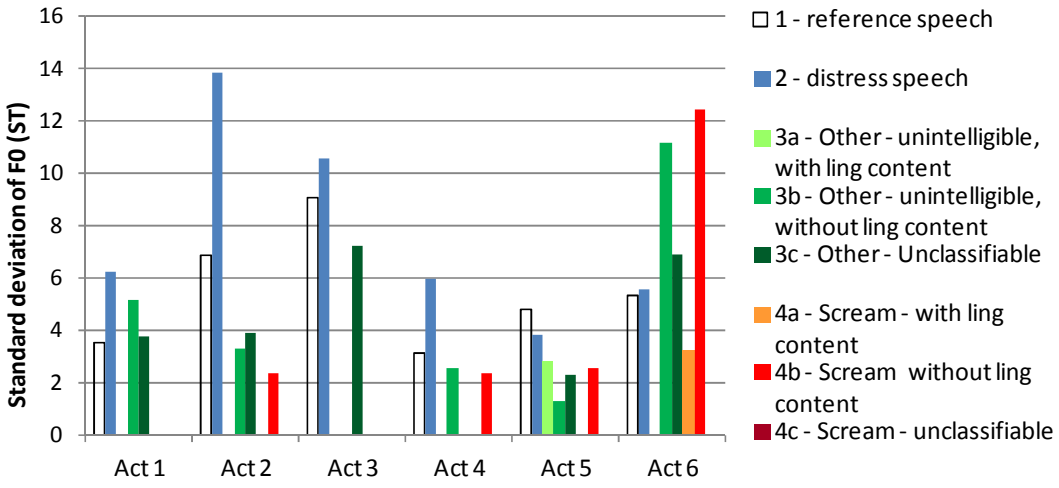
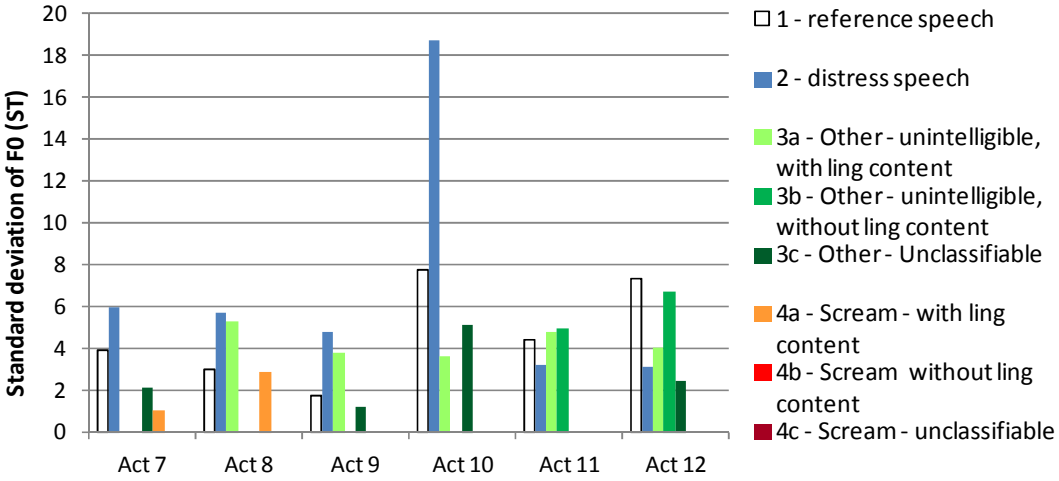


Figure 5-16: Standard deviation of F0 (semitones) for each category of the distress taxonomy in female actors.

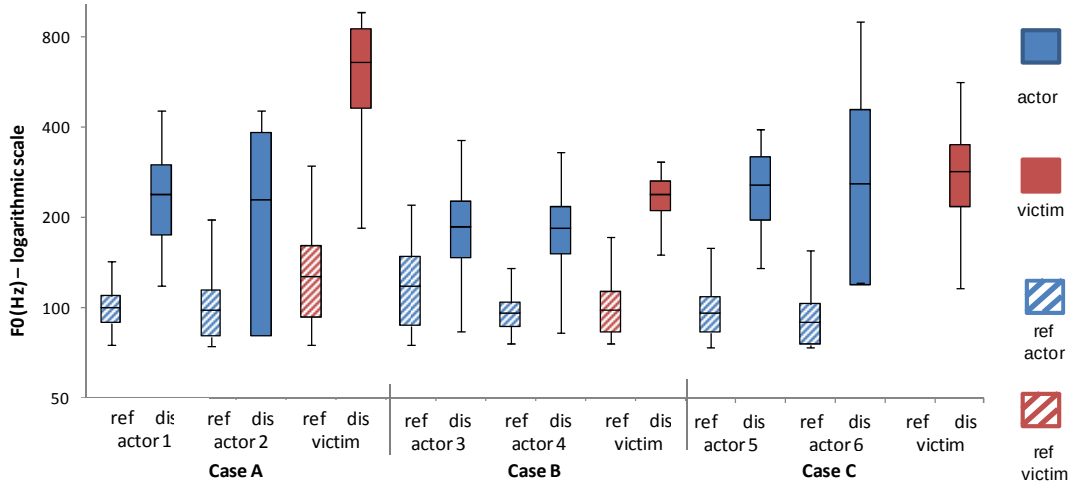


5.1.3 Victim vs. actor distress

Displayed below are boxplots amalgamating F0 data from actors' and victims' reference and distress material from the previous two sections.[15] Figure 5-17 compares male victims and actors from Cases A, B, C, and Figure 5-18 compares female victims and actors from Cases C, D, and E.

[15] For actors, distress material here refers to their rehearsed performance.

Figure 5-17: F0 mean, S.D., min. and max. for male actors and victims.



In Figure 5-17, the mean F0 of reference speech falls in the typical male range (from 90 to 130 Hz). In distress speech, the mean F0 for male actors typically increases to 200-300 Hz, with male victims showing a mean F0 greater than that of actors. Some are within the same range, i.e. 200-300 Hz, but some are much greater than others, e.g. Victim A exhibits a mean of 659 Hz. Inter- and intra-speaker variation in F0 mean and range is evident. Victim A has the highest F0 at 968 Hz, though actor 7 also exceeds 900 Hz; therefore, actors are capable of physically producing as high an F0 as real victims. Victim C's highest value is 563 Hz, yet most male actors, and also victim B, do not exceed 500 Hz. In these three examples, the F0 was realised when the actor or victim was screaming, not during speech.

Figure 5-18: F0 mean, S.D., min. and max. of female actors and victims.

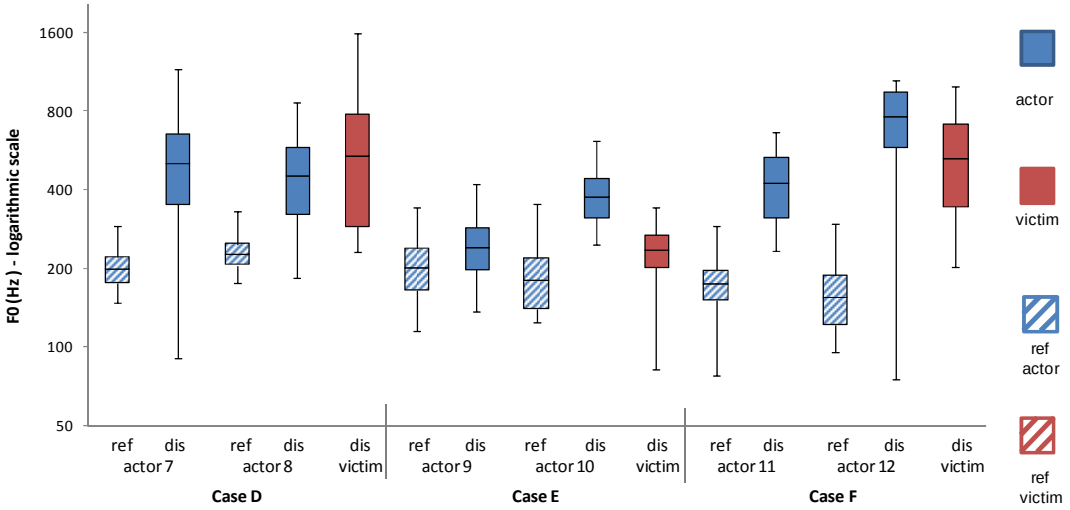


Figure 5-18 paints a similar picture for the female actors and victims. The mean F0 values drawn from reference material for the female actors typically falls between 170 and 210 Hz – in line with average values for adult females from clinical phonetic research (e.g. Baken (1987)). F0 in distress conditions increases, with means ranging from 242 Hz to 767 Hz. Unlike the male actors, who tend not to increase in F0 as much as the male victims, the female actors appear to be closer to producing similar F0s to those of the victims, and three actors (actors 7, 10 and 12) even exceed the mean distress F0 of the corresponding victim. Inter- and intra-speaker variation in F0 mean and range are, again, noticeable, with Victim D reaching 1580 Hz when screaming. Actors 7 and 12 also reach over 1000Hz and Victim F reached 995 Hz. In a similar fashion to that of the male actors and victims, these peak values are achieved in screamed productions. There is no statistically significant difference for the actors with regard to mean F0 values and standard deviation in reference versus distress conditions. This is also true for the victims.

Focussing on the mean F0 data, and adding unrehearsed F0 material to the comparison, Figure 5-19 reiterates the finding that in most cases actors' rehearsed and unrehearsed material tends to be similar, especially among males. The inter-speaker variation in F0 mean is also very apparent, as is further illustrated in Figure 5-21, where the increase in semitones from reference to distress material ranges from 3.2 ST for Actor 9 to 28.5 ST for Victim A. On the whole, most speakers produced

an increase in F0 of around 10-15 ST in distress conditions. Again, no statistically significant change is observed with respect to the increase in F0 measured in semitones for either actors or victims.

Figure 5-19: Mean F0 for reference, unrehearsed distress and rehearsed/real distress across male actors and victims.

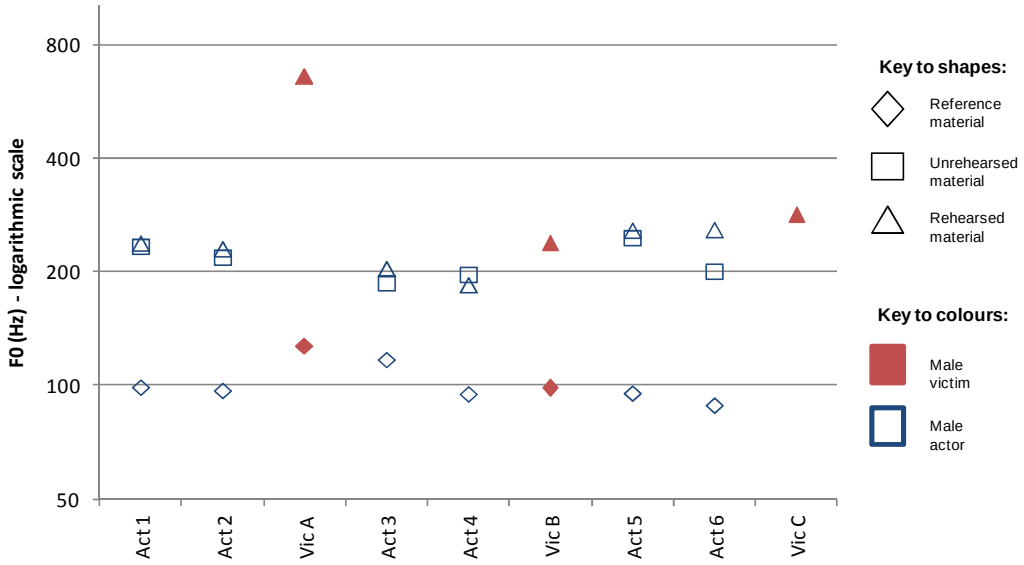


Figure 5-20: Mean F0 for reference, unrehearsed distress and rehearsed/real distress across female actors and victims

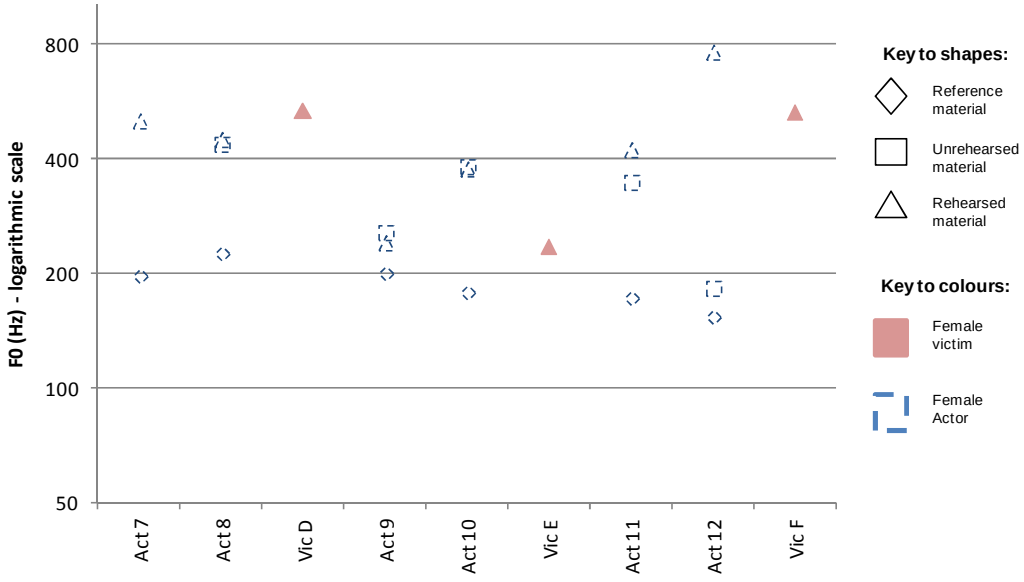
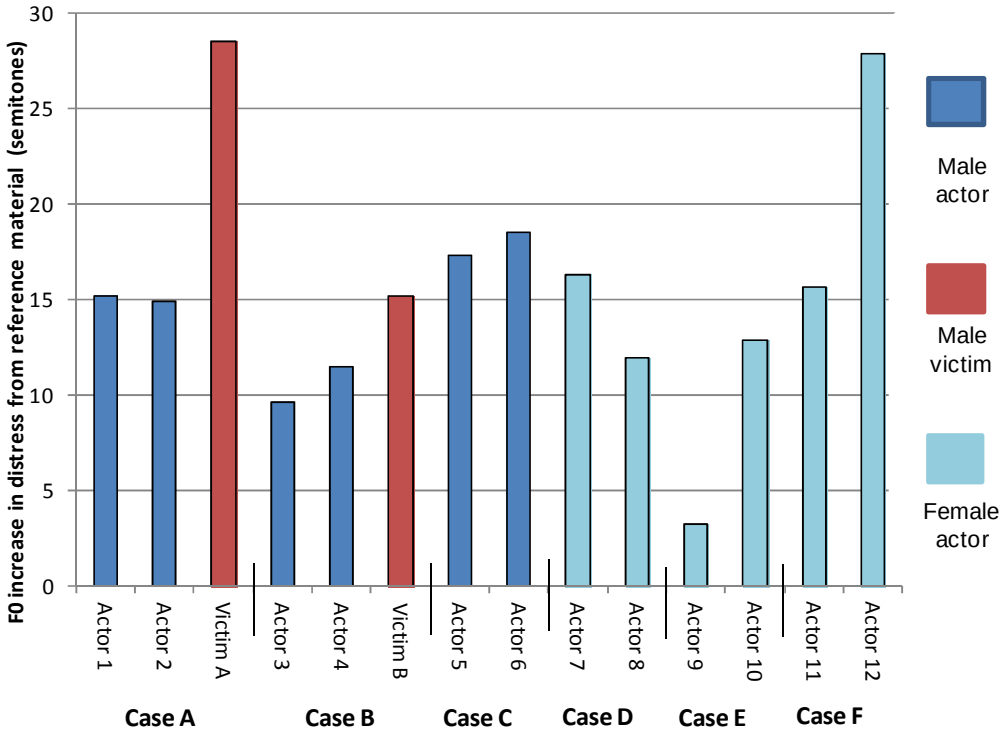


Figure 5-21: Increase in semitones from reference material to rehearsed (actor) or real (victim) distress material.



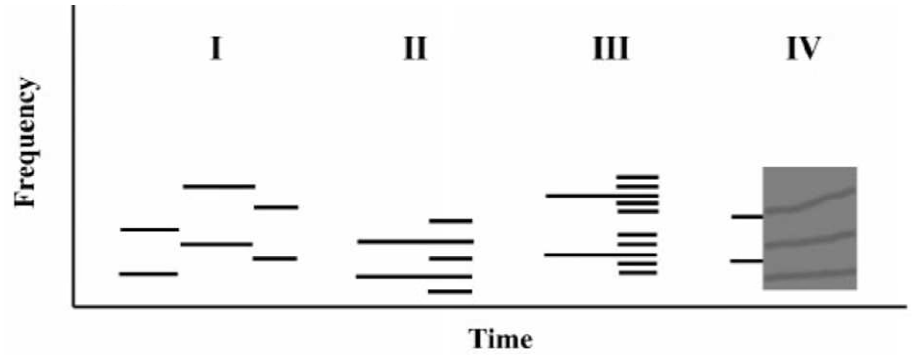
In comparing the manner of distress productions between actors and victims, there were no statistically significant changes between the proportion of each category when analysing categories of distress speech and other vocalisations (categories 2 and 3). However, there was a statistically significant difference between the proportion of screamed categories (category 4) in actors and victims, with victims producing screams in distress more than actors (Mann-Whitney $U = 53.0$, $z = 1.701$, one-tailed, $p < 0.045$, $r = 0.4$). There was also a significant change in the standard deviation of screamed productions, with victims typically having a higher F0 standard deviation when screaming than actors (Mann-Whitney $U = 22.0$, $z = 2.132$, two-tailed, $p < 0.033$, $r = 0.67$). With the exception of a significant difference between actors' and victims' standard deviations of distress speech (category 2), where Mann-Whitney $U = 10.00$, $z = -1.698$, two-tailed, $p < 0.045$, $r = -0.42$, there were no further statistically significant changes in F0 mean and S.D. across categories.

5.1.4 Miscellaneous observations

During the acoustic analysis, an additional characteristic of distress data in the form of the presence of non-linear phenomena was observed in the spectrograms. Using

the definitions from Riede et al.(2004), bifurcations such as F0 jumps (sudden changes in F0 due to an abrupt and discontinuous increase or decrease in vocal fold vibration), subharmonics (frequencies that occur as a fraction of the F0, e.g. F0/2, F0/3 etc, visible as additional spectral components), biphonation (the presence of two independent F0s as separate frequency contours in the spectrogram), and deterministic chaos (episodic non-random noise which is typically characterised by proximity to subharmonics and abrupt onsets/offset), were identified in both the acted and authentic data. Figure 5-22 shows schematic narrowband spectrograms of these non-linear phenomena.

Figure 5-22: Schematic narrowband spectrograms illustrating non-linear phenomena: frequency jumps (I), subharmonics (II), biphonation (III), and deterministic chaos (IV) from Riede et al. (2004: 278, their Figure 1b)



These bifurcations are examples of sound productions using non-linear source-filter coupling. The traditional source-filter theory of speech production assumes that the source and filter act independently of each other (Fant 1971), and it has been used successfully to describe acoustic features of vowels and voiced consonants over the past 50 years. Non-linear coupling differs from linear source-filter coupling primarily in terms of the source impedance in relation to the filter impedance. As described in Titze (2008), for linear speech production, the firm adduction of the vocal folds and the widening of the vocal tract area within the larynx (the epilaryngeal tube) results in the source impedance (transglottal pressure divided by glottal flow) being higher than the filter impedance (vocal tract input pressure divided by the airflow into the vocal tract). The width of the epilaryngeal tube ensures a mismatch between the two impedances. In non-linear sound production, the source and filter impedances are comparable due to a narrow epilaryngeal tube

coupled with vocal fold adduction. The pressures in the vocal tract contribute to the production of frequencies at the glottis, leading to bifurcations.

In the acted and authentic distress, examples of non-linear productions were found, indicating that when producing distress productions, some individuals will narrow their epilarynx (presumably due to increased muscle tension).

Figure 5-23: Waveform and narrowband spectrogram illustrating both linear and non-linear sound production in a scream produced by Actor 5.

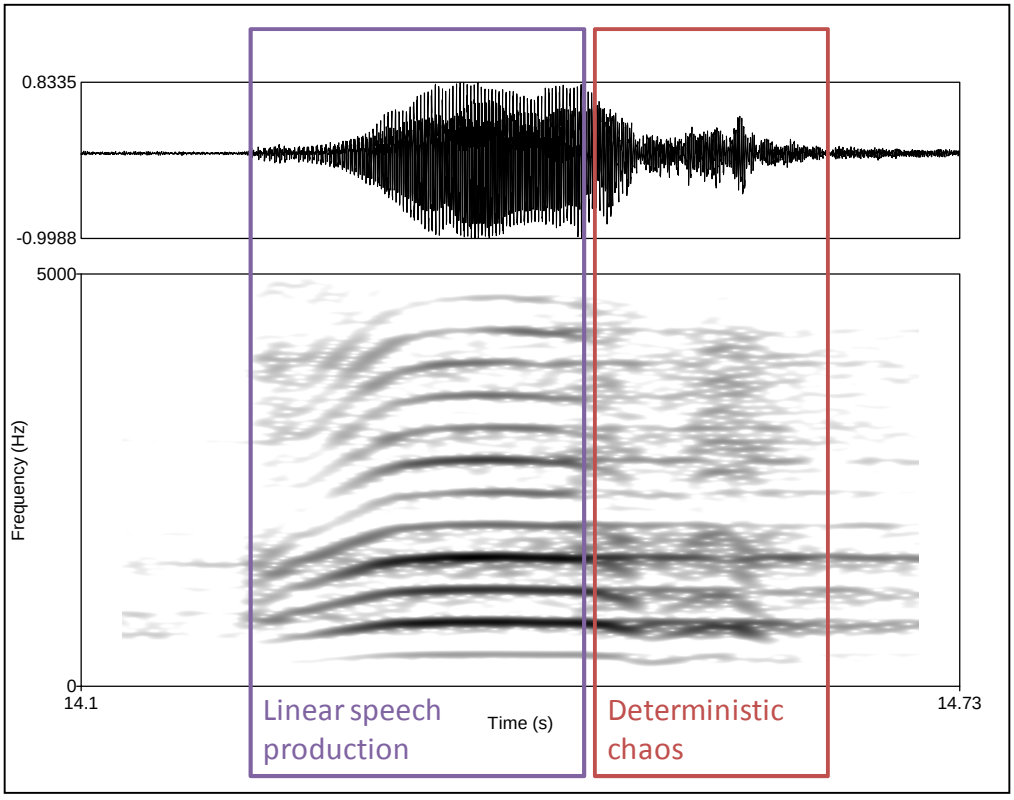


Figure 5-23 shows a screamed production by Actor 5 which is characterised by both linear and non-linear sound production. The start of the scream shows a smooth increase in F0 with clearly defined harmonics. In the end portion of the scream, the harmonics are less clearly defined and punctuated with brief periods of broadband noise with a sudden onset. A screamed production by Victim C also contains bifurcations, but in his case he exhibits F0 jumps and subharmonics. In contrast to Actor 5, the majority of Victim C's bifurcations occur in the first half of his screamed production.

Figure 5-24: Narrowband spectrogram and waveform illustrating F0 jumps and subharmonics in a vocalisation produced by Victim C.

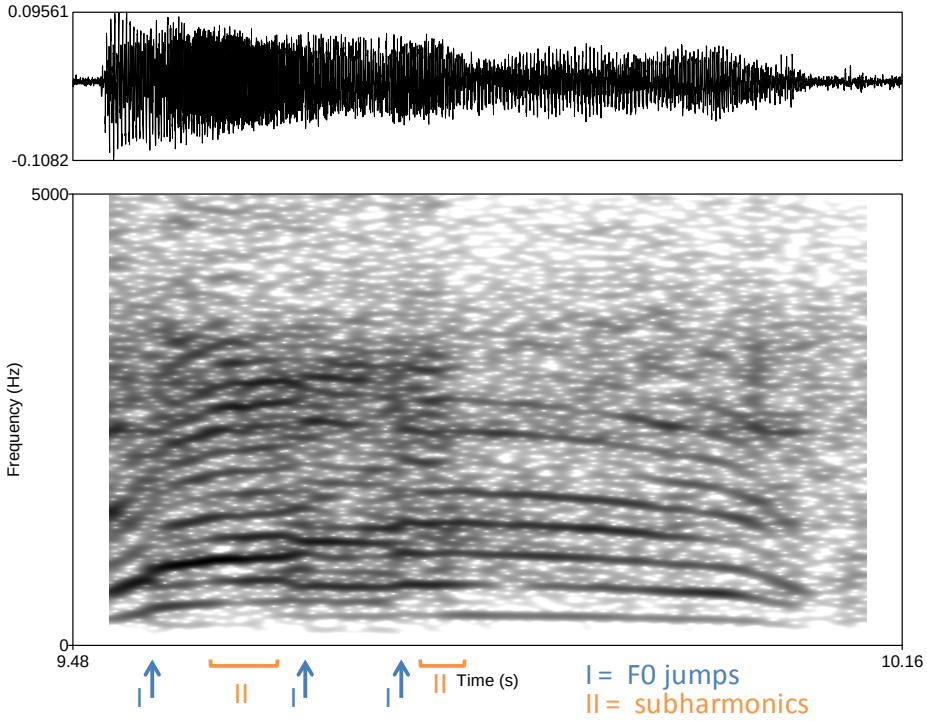
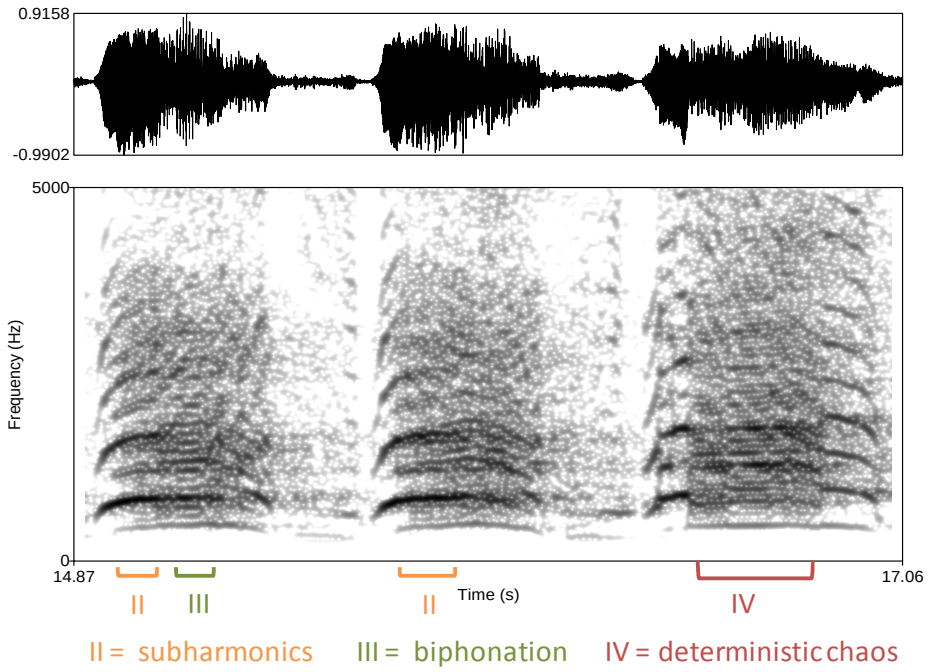


Figure 5-25: Narrowband spectrogram and waveform illustrating subharmonics, biphonation and deterministic chaos in a series of screams produced by Actor 6.

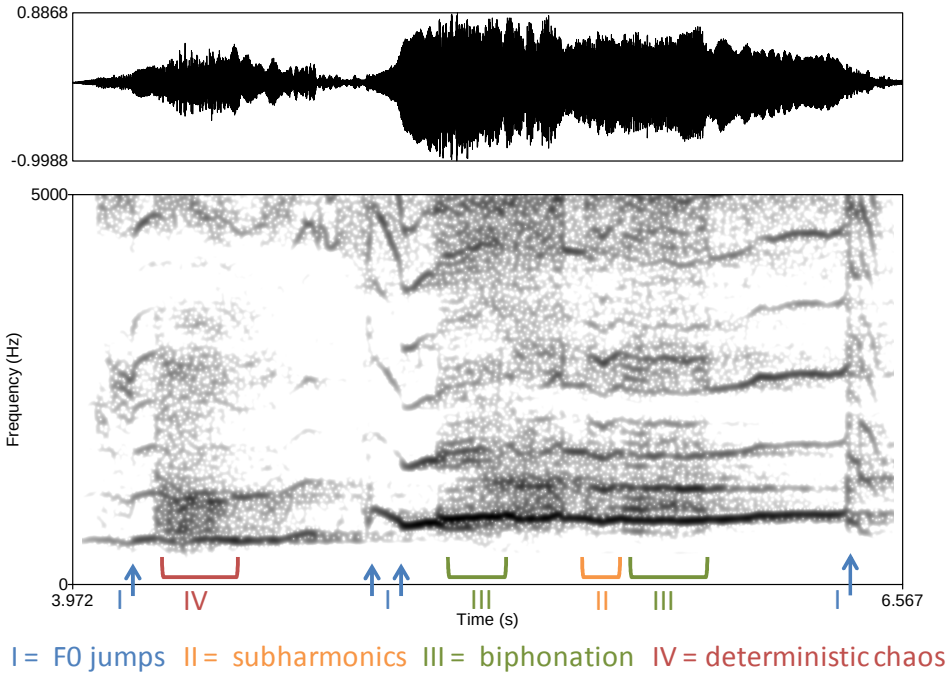


A combination of subharmonics, biphonation and deterministic chaos are present in a series of screams produced by Actor 6 (Figure 5-25). In all three of the previous

examples, male individuals producing either vocalisations or screams exhibit non-linear phenomena. These productions all feature a high F0. Not all high-F0 productions result in bifurcations, but non-linear phenomena are not observed in any reference or intelligible speech produced by the actors or the victims.

The bifurcations in the distress data are not just limited to male actors and victims. An 'other'-style vocalisation produced by (female) Actor 12 contains all four types of bifurcation (Figure 5-26). Akin to the previous examples, this production was also produced with a high F0.

Figure 5-26: Narrowband spectrogram and waveform illustrating four types of bifurcation in a high-F0 vocalisation produced by Actor 12.



Inspecting the distress data for non-linear phenomena was not part of the planned acoustic analysis and therefore they were not measured in a quantifiable way. However, the presence of bifurcations in high-F0 productions in both male and female speakers, and in both victims and actors, is an interesting finding. Although non-linear phenomena are well-reported in animal (mammalian) literature and disordered speech literature (see §8.1.1for further discussion), this is the first recorded observation in distress speech data.

5.1.5 Summary of F0 results

Findings from the analyses of F0 in actors and victims can be summarised as follows:

- Both actors and victims show considerable inter- and intra-speaker variability in F0 production.
- F0 increases from reference to distress material in victims (not tested statistically).
- F0 increases in actors' distress means and standard deviations relative to their reference speech are statistically significant
- There are no statistically significant changes between actors' rehearsed and unrehearsed distress conditions with respect to F0 mean, S.D., or increase in semitones.
- There are no statistically significant differences between male actors' and victims' F0 data, nor are there any statistically significant differences between female actors' and victims' data.
- There are no statistically significant differences in the proportions of categorisations of speakers' vocal responses (based on the distress taxonomy) between actors' rehearsed and unrehearsed conditions, nor between male and female actors.
- Statistically significant differences are observed between actors and victims with respect to their proportion of screamed productions, their standard deviation of screamed productions, and their standard deviation of distress speech.
- Non-linear phenomena such as various types of bifurcation are present in several high-F0 productions in both male and female actors and actresses.

We can conclude that observable, statistically significant differences are apparent in F0 data between reference and distress conditions, though fewer significant differences are found between actors and victims. Acoustic differences between actors' and victims' distress productions are only apparent when F0 is analysed using a categorisation system that compares similar manners of distress speech within each speaker, and the change is mainly observed when comparing actors' and victims' screamed productions.

5.2 Intensity

The following four sections of this chapter report on the findings for acoustic intensity, measured in decibels (dB). The intensity findings are expressed in both absolute and relative forms. Variation in intensity amongst the actors forms the focus of this section by virtue of the greater control exercised when collecting the acted data. Variation amongst the victims is also assessed and commented on, but without more controlled data (which would be difficult to acquire given the nature of forensic data) observations are, at this stage, qualitative.

5.2.1 Reference vs. distress in victims

Intensity measurements for victims must be treated with caution as it is impossible to account for changes in distance from, and orientation of, the speaker to the microphone of the recording device. However, intensity was analysed in Praat and, although the absolute mean, minimum, maximum results are without meaning given the different recording environments for each speech condition, the relative intensity changes across speech conditions may be useful. Table 5-4 shows intensity data for victims A and B in reference and distress conditions. Both show a similar pattern in that the standard deviation of intensity of their productions increases in distress (from 6.6 dB to 12.4 dB for Victim A, and 7.0 dB to 13.4 dB for Victim B). Both show similar levels of standard deviation in the respective conditions, approximately7 dB for reference material, and 13 dB for distress. The remaining victims each show a standard deviation range between 5.2 and 9.0, but without reference data it is impossible to know whether this represents an increase, decrease or an equivalent level in standard deviation.

Table 5-4: Intensity mean, S.D., min., max., and for all victims in reference (where possible) and distress conditions.

Speaker	Mean (dB)	S.D. (dB)	Min. (dB)	Max. (dB)
Vic A - ref	68.11	6.58	48.85	86.18
Vic A	70.83	12.38	48.83	86.11
Vic B - ref	68.36	7.03	53.31	83.61
Vic B	62.21	13.44	27.33	81.98
Vic C	56.97	7.92	37.79	69.75
Vic D	73.89	6.16	46.46	85.92
Vic E	78.67	5.20	65.45	86.09
Vic F	70.62	9.01	51.22	83.79

If we look at changes in standard deviation after having categorised victims' distress responses as per the taxonomy (and averaging all 'other' and 'scream' responses together, i.e. 3a-c considered as one category, and 4a-c forming one category), we see that, with the exception of Victim B, all 'other' responses (i.e. those considered unintelligible with varying degrees of perceptible linguistic content) tend to be more variable, i.e. have a higher standard deviation, than intelligible distress speech, and screamed productions tend to have the lowest variability (Table 5-5).

Table 5-5: Intensity standard deviation across different categories of distress response for all victims.

Speaker	All distress (aggregate average)	Cat 2 - speech	Cat 3 - other	Cat 4 - scream
Vic A	12.38	x	8.38	3.53
Vic B	13.44	12.32	8.83	x
Vic C	7.92	7.60	7.77	6.96
Vic D	6.16	5.15	6.22	3.63
Vic E	5.20	4.55	x	x
Vic F	9.01	x	9.59	6.57

An open question is whether we see a change of standard deviation variability according to whether there is perceptible linguistic content in sections categorised as 'other' or 'scream'. Unfortunately, none of the victims' productions was categorised as '3b - other vocalisation without linguistic content' so no comparison could be made, and only two victims (C and D) had screamed material including both 'with linguistic content' and 'without linguistic content' classifications. In both cases, the standard deviation was higher in the 'without linguistic content' condition. Victim C's screamed productions increased from 5.94 (with linguistic content) to 7.99 (without linguistic content), and Victim D increased in the same way from 3.58 to 4.02 (Appendix F2b).

When inspecting the victims' high F0 distress material as sound waves and spectrograms, the intensity contour tends to follow a similar pattern to the pitch contour (e.g. Victim C in Figure 5-27 and Victim F in Figure 5-28).

Figure 5-27: Waveform with F0 and intensity contours for Victim C showing a screamed production without linguistic content (first half of waveform) and an 'other' vocalisation that had unclassifiable linguistic content (second half of waveform).

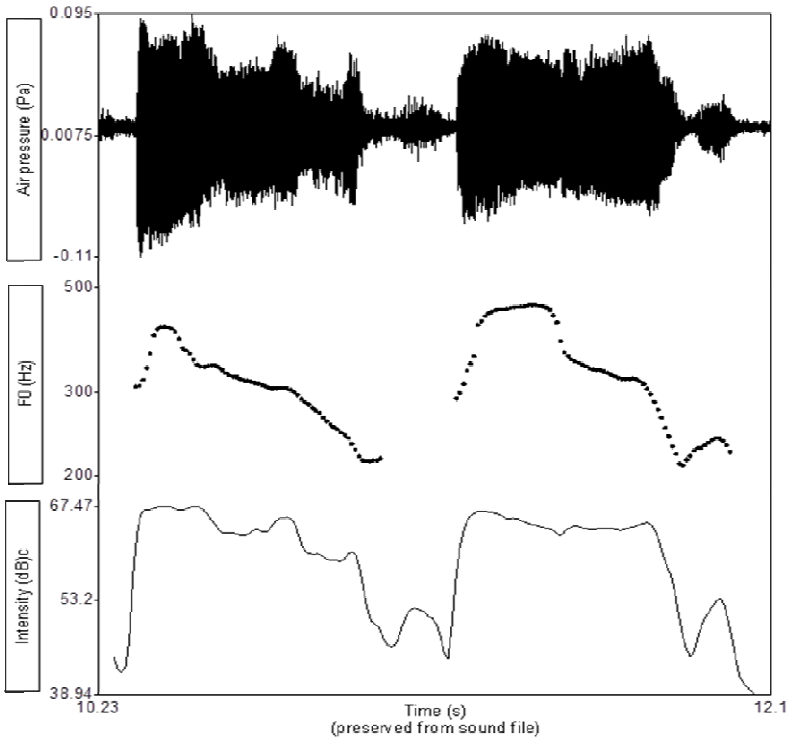
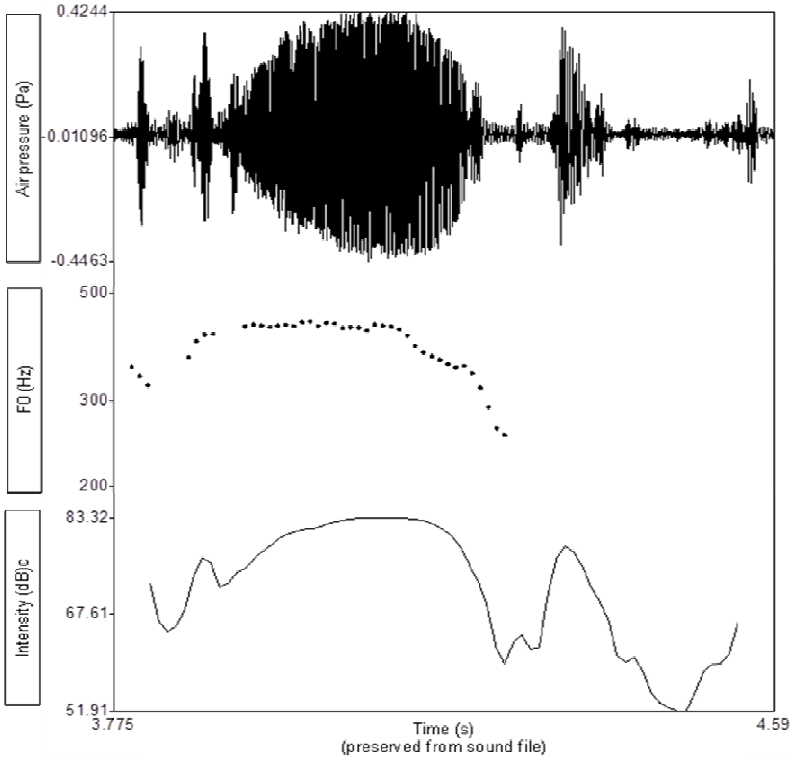


Figure 5-28: Waveform with F0 and intensity contours for Victim F producing an 'other' vocalisation that had unclassifiable linguistic content.



5.2.2 Reference vs. distress in actors

Since actors wore headband microphones that remained stable at a fixed point from their mouths in all speech conditions, comparisons between their mean intensity levels across conditions can be made. Figure 5-29 and Figure 5-30 illustrate absolute intensity mean, minimum, maximum and standard deviation in male and female actors in all three speech conditions using modified boxplots. In both sets of actors a decrease in mean intensity in distress speech conditions, especially rehearsed, and an increase in standard deviation, also in rehearsed distress speech, can be seen. Both patterns are more marked among the female actors. Minimum intensity levels in distress fluctuate quite markedly across actors, yet maximum levels appear to peak between 85dB and 90dB for everyone. This plateau in intensity may be the consequence of limitations of the equipment.

Figure 5-29: Mean, min., max., and S.D. of intensity (dB) in male actors in all speech conditions.

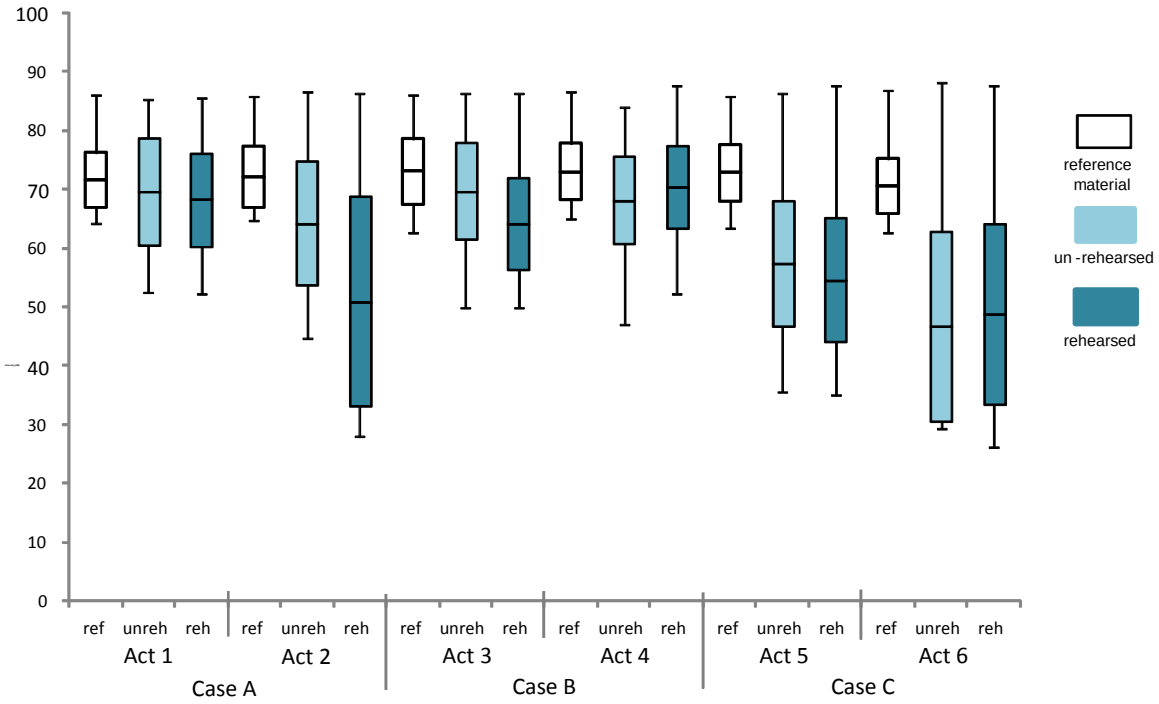
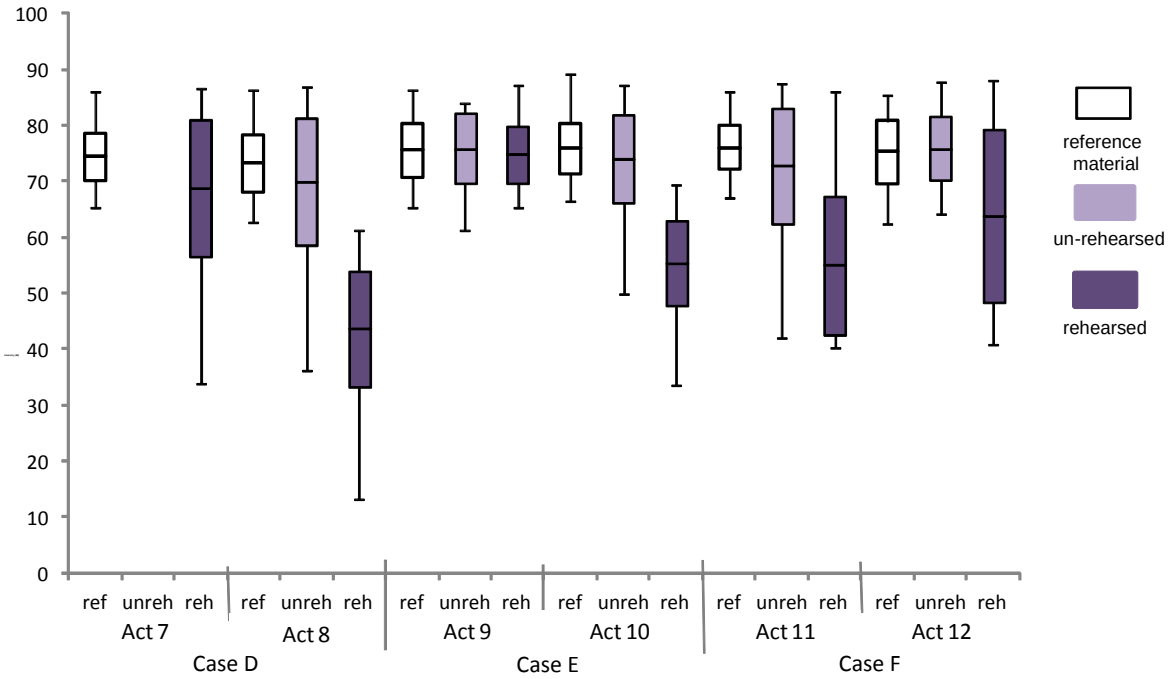


Figure 5-30: Mean, min., max., and S.D. of intensity (dB) in female actors in all speech conditions.



A repeated measures one-way ANOVA showed a significant effect of a decrease in intensity for the three conditions (F = 14.55, df = 2, p <0.001). The means of intensity in all three speech conditions (reference, unrehearsed and rehearsed) were 73.53 (S.D. = 1.87), 67.67 (S.D. = 8.78) and 59.09 (S.D. = 9.87), respectively. Therefore, actors decreased in intensity from reference to unrehearsed distress speech, and then decreased further in rehearsed distress speech. Three related *t*-tests were performed (with a Bonferroni correction, i.e. effects are reported at a 0.0167 level of significance). These showed that the difference between reference and rehearsed distress speech, and the difference between rehearsed and unrehearsed distress speech, are significant (t = 4.99, df = 11, p < 0.001; t = 2.99, df = 10, p = 0.14 respectively). There is no statistically significant difference between reference and unrehearsed speech.

The column charts below show relative differences, rather than absolute values, in intensity scores in actors' reference and distress performance recordings. A negative difference value shows a decrease of intensity in the distress condition, and a positive score demonstrates an increase in intensity in the distress condition. Figure 5-31 illustrates that some male actors, e.g. actors 1, 2, and 6, increased their mean intensity in distress performances, whereas others decreased it, e.g. actors 3, 4, and 5.

In most cases, the direction of the change was uniform in both rehearsed and unrehearsed distress performances, though exceptions are actors 1 and 6, with both showing an increase in mean intensity during their unrehearsed performance, but an almost negligible decrease in their rehearsed performance. On the whole, the trend is for male actors to decrease mean intensity in rehearsed distress material.

Figure 5-31: Differences in mean intensity among male actors in rehearsed and unrehearsed distress from reference material.

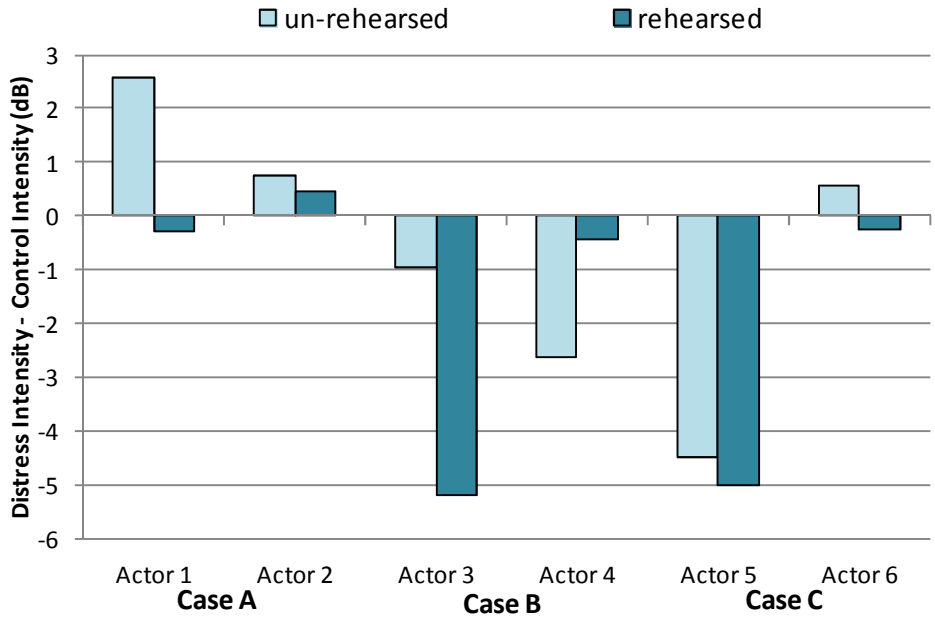
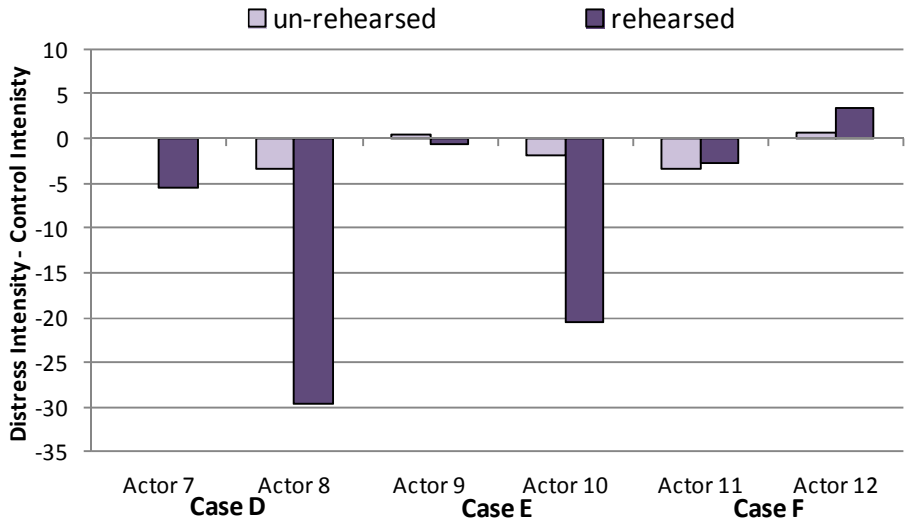


Figure 5-32 on the following page shows that most female actors decreased in mean intensity when performing distress (actors 7, 8, 10 and 11), with only one actor (actor 12) showing a small increase in mean intensity in rehearsed and unrehearsed conditions. Actor 9 showed negligible differences in intensity between her distress and reference material. Akin to the F0 results, the extent of the increase or decrease of mean intensity in distress performances varied across actors.

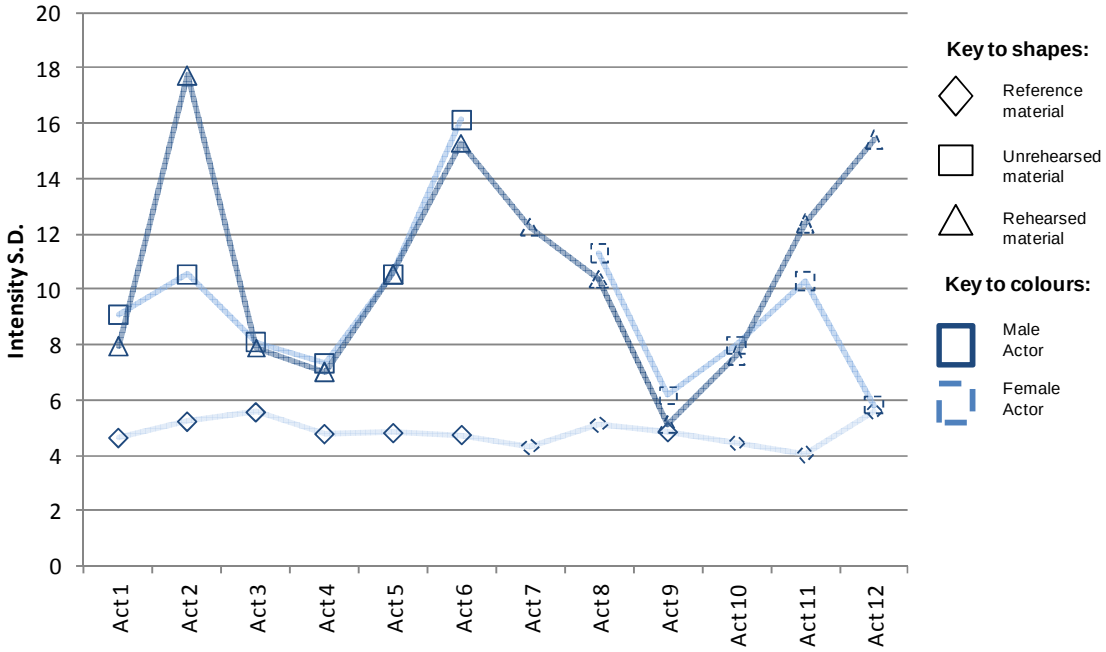
There were no statistically significant differences between the relative intensity differences for all actors in rehearsed and unrehearsed conditions, nor between male and female actors in both of these distress speech conditions.

Figure 5-32: Differences in mean intensity among female actors in rehearsed and unrehearsed distress from reference material.



As first illustrated in Figure 5-29 and Figure 5-30, there appears to be a tendency for the standard deviation of intensity to increase in distress conditions. Figure 5-33 displays the average level of standard deviation by all actors across the three speech conditions and demonstrates the extent of the increase from reference to unrehearsed and rehearsed conditions. For some actors (Actors 3, 4, 5, 6, 8, and 10), the increase in standard deviation in rehearsed and unrehearsed conditions is quite similar, whereas in others (Actors 2, 11 and 12) a more dramatic increase between unrehearsed and rehearsed distress speech can be seen.

Figure 5-33: Levels of standard deviation for intensity across all speech conditions for all actors.

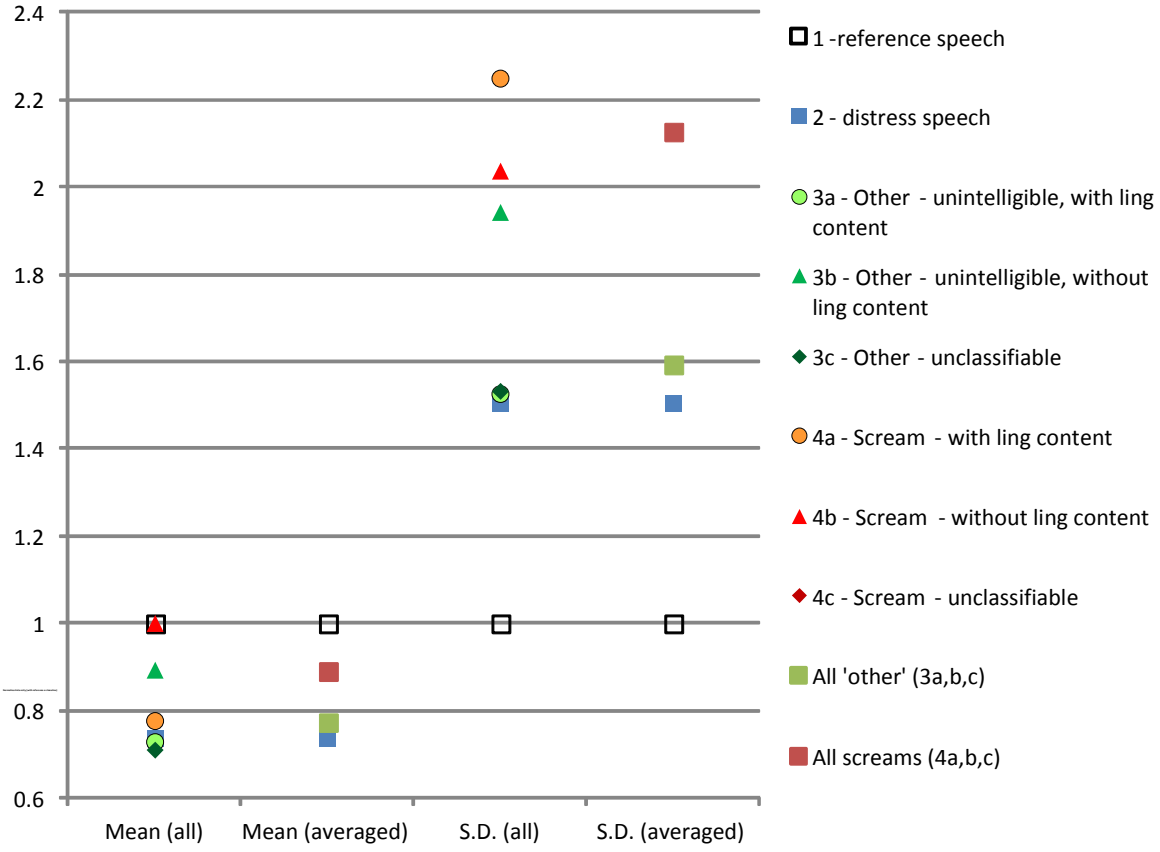


A repeated measures one-way ANOVA showed a significant effect for changes in standard deviation across the three speech conditions (F = 15.73, df = 2, $p < 0.001$). The means of standard deviation in all three speech conditions were 5.89 (S.D. = 0.47), 9.39 (S.D. = 2.89) and 10.66 (S.D.= 4.07) respectively. Consequently, we can say actors' standard deviations of intensity levels increased from reference to unrehearsed distress speech, and then increased further in rehearsed distress speech. Three related *t*-tests were performed (with a Bonferroni correction applied so that effects are reported at a 0.0167 level of significance); these showed that the difference between reference and unrehearsed distress speech, and the difference between reference and rehearsed distress speech, are significant (t = -4.90, df = 10, p = 0.001; t = -5.35, df = 11, $p < 0.001$, respectively). There was no significant difference between the standard deviation in actors' rehearsed and unrehearsed speech conditions.

In order to see whether intensity mean and standard deviation are affected by the manner of response, i.e. as categorised using the distress taxonomy, normalised intensity mean and standard deviation across all actors and categories in reference

and rehearsed distress speech were calculated. These are displayed in Figure 5-34.[16] It shows the mean and standard deviation of intensity across all taxonomy categories and subcategories, as well as averaged values across the four principal categories. It, too, exemplifies the decrease in mean intensity from reference to distress speech material, but it also highlights that amongst just the distress categories, screams have a higher intensity than 'other', which in turn has a higher intensity than intelligible distress speech. Standard deviation shows a similar pattern in that screamed productions have a higher S.D. than 'other' productions, which again have a higher S.D. than distress speech. However, unlike the intensity means, the intensity standard deviations feature increases that are higher than the reference speech S.D.

Figure 5-34: Normalised intensity mean and standard deviation across all categories by all actors.



A repeated measures one-way ANOVA showed a significant effect for mean intensity levels in categories of distress production (F = 16.52, df = 2, p = 0.001).

[16] The average mean and S.D. of each category across all speakers were first calculated and then normalised using the formula 1/(A/B) where A = averaged reference value and B = averaged category value.

The means of intensity in intelligible distress speech, averaged 'other' productions, and averaged screamed productions, were 54.16 (S.D. = 13.84), 56.90 (S.D. = 12.84) and 65.53 (S.D. = 13.43) respectively. Therefore, the mean intensity increased from distress speech to 'other' productions, and then increased further in screamed productions. The difference between distress speech and screamed productions, and the difference between 'other' productions and screamed productions, were found to be significant (t = -4.60, df = 4, p = 0.010; t = -4.92, df = 4, p = 0.08, respectively) after having conducted three related t-tests (with Bonferroni correction so that effects are reported at a 0.0167 level of significance). There is no statistically significant difference between distress speech and 'other' productions. Similarly, for standard deviation of intensity across distress categories, a repeated measures one-way ANOVA showed a significant effect (F = 10.79, df = 2, p = 0.005). The means of standard deviation in the three distress speech categories were 7.28 (S.D. = 1.06), 7.71 (S.D. = 1.92) and 10.29 (S.D. = 1.12) respectively. Consequently, we note that standard deviations of intensity levels increased from distress speech to 'other' productions, and then increased further in screamed productions. Applying a Bonferroni correction (with effects reported at a 0.0167 level of significance), three related t-tests were performed, which showed the difference between distress speech and screamed productions to be significant (t = -8.64, df = 4, p = 0.001). There is no significant difference between the standard deviations of distress speech and 'other' productions, nor between 'other' productions and screams.

5.2.3 Victim vs. actor distress

In §5.2.1, it was acknowledged that absolute intensity scores for victims were meaningless at this stage due to the uncontrolled recording environment. However, three patterns were observed that can be compared with the acted data set. Firstly, it was noted that in the victims' productions of high F0, the intensity contour tended to mimic the pitch contour. Since §5.1 demonstrated significant increases in F0 in distress productions, we can hypothesize an increase in intensity from reference to distress speech conditions. However, as reported in the previous section and illustrated from Figure 5-29 through Figure 5-33, intensity variation across speech conditions does not confirm the predicted pattern. Instead, the opposite is found, i.e. a significant decrease in mean intensity levels is observed from reference to distress

conditions. Within the distress speech productions of the actors, there are significant differences between productions classified as intelligible speech, 'other' and screamed, with screams featuring a significant increase in mean intensity as compared with the other two distress conditions. Figure 5-34 highlighted this hierarchy and illustrated that when acted distress data were not analysed with regard to the corresponding reference material, the mean intensity of each category increased in the same order that was observed in the acted F0 data (Figure 5-13 and Figure 5-14).

Secondly, victim data which allowed for a comparison between reference and distress material, which was limited in quantity, suggested an increase in the standard deviation in distress speech conditions. This was proven to be statistically significant in the acted data. However, there was no statistically significant difference between the (aggregate) distress S.D. for actors versus victims.

Thirdly, examining the standard deviation of intensity using categories of distress highlighted a possible difference between actors and victims, in that for actors, screamed productions tend to have the lowest distress S.D. of the three distress categories, with 'other' productions, i.e. unintelligible vocalisations, having the highest S.D. This pattern was not borne out in the acted data, where significant changes were observed across distress categories but in a different direction; intelligible distress speech had the lowest S.D. followed by 'other', and then screams had the highest S.D. (echoing the same hierarchy as the categorised mean intensity scores for actors). The average S.D. in all screamed productions of actors (mean = 10.3, S.D. = 1.12) is significantly higher than that for the victims (mean = 5.18 , S.D. = 1.85) (t = 5.18 , df = 7 , two-tailed p = 0.001). This shows that screamed productions by actors were significantly more variable in terms of their intensity than those produced by victims. There were no significant changes between actors and victims in any other distress categories.

5.2.4 Summary of intensity results

§5.2 examined differences in acoustic intensity between reference and distress speech, and between actors' and victims' distress responses. The difficulties in measuring intensity in authentic material, discussed in §3.3.2.2, mean that

meaningful observations about intensity variation in genuine material are limited. Victims show much inter- and intra-speaker variability in intensity, though this variability is no doubt complicated by the lack of controlled environment. There is also variation within and across actors' intensity levels, though not to the same degree as for victims, and in fact a plateau effect is noted for maximum intensity levels, though this may in part be due to the constraints of the recording equipment, rather than physiological limitations. The principal findings are summarised as follows:

- Significant decreases are observed in mean intensity from reference to distress material in actors.
- A decrease in intensity from reference to distress material is not observed among victims. (In fact, we might expect the opposite, given observations of increased intensity in high-F0 productions in victim data).
- For actors there is a significant increase in intensity across the different distress categories; screamed productions tend to have a higher mean intensity than 'other'-labelled vocalisations, which in turn have a higher mean intensity than distress speech.
- An increase in intensity standard deviation between reference and distress conditions is apparent for both actors and victims, and significantly so for the actors.
- There was no significant difference between actors' and victims' increase in intensity standard deviation.
- Intensity S.D. significantly increased from intelligible distress speech to unintelligible 'other' vocalisations to screamed productions among actors (echoing the pattern of the increase in mean intensity across the distress taxonomy by the actors).
- In the authentic victim data, unintelligible 'other' productions had the highest standard deviation for intensity, and screams had the lowest (though the paucity of observations should be taken into account).
- A significant difference was found between actors and victims in terms of standard deviation of their screamed productions. Actors were considerably more variable when screaming than were the victims.

We can conclude that, as with F0, there are some observable and significant differences in intensity that can differentiate reference from distress speech produced by actors. Specifically, intensity decreases from reference speech to distress productions. When comparing acted and authentic screams, we can differentiate actors from victims on account of the actors' significant increase in intensity variability. However, it should be noted that this finding may not be generalisable beyond the current data set.

5.3 Articulation rate

The articulation rates presented below are based on the methodology described in §3.3.2.3. The articulation rates are calculated by counting the number of phonetic syllables produced per second, excluding pauses of over 100 ms (Künzel 1997). The amount of speech material available for this analysis varied from 3s to 50s (Table 5-6). For some distress data, very little speech material was available due to the nature of the original distressing context, e.g. Case A, where the victim reacts to a gun being drawn and fired, and utters four words in between shots. Victim A produces just over 3s of speech material, and similar amounts of material are replicated by actors re-enacting the scenario. Victim D, on the other hand, produces approximately 36s of speech material while fleeing her attacker and simultaneously talking to the operator. The actors re-enacting this case produced between 48s and 55s of speech. Reference speech material produced by Victim A resulted in 18s of material, and reference material from Victim B was 22s in length. Reference speech material produced by the actors ranged between 40s and 50s in duration.

Although the AR is calculated for even the shorter extracts, the utility of such a measurement is not beyond question. In some cases, specifically for the brief extracts of speech material which contain screams and 'other'-type vocalisations, the AR measurement reflects a duration measurement of that category of speech, rather than the rate of articulation. As described in the following section, there might still be some value in such measurements, but caution should be exercised when interpreting these data.

Table 5-6: The number of phonetic syllables and the duration of speech samples (excluding pauses) produced by victims and actors across all speech conditions.

			No. of phonetic sylls	Sample duration (s)
Case A	Vic A	reference	140	22.06
		distress	7	4.77
	Act 1	reference	236	49.46
		unrehearsed	10	3.02
		rehearsed	8	4.56
	Act 2	reference	234	44.25
		unrehearsed	9	3.53
		rehearsed	9	3.86
Case B	Vic B	reference	96	18.39
		distress	101	23.19
	Act 3	reference	232	41.81
		unrehearsed	120	26.19
		rehearsed	110	24.50
	Act 4	reference	231	44.01
		unrehearsed	128	27.34
		rehearsed	121	28.03
Case C	Vic C	reference	NA	NA
		distress	48	14.38
	Act 5	reference	233	43.66
		unrehearsed	67	17.35
		rehearsed	66	17.02
	Act 6	reference	231	46.90
		unrehearsed	68	17.29
		rehearsed	67	19.14
Case D	Vic D	reference	NA	NA
		distress	94	35.87
	Act 7	reference	232	40.95
		unrehearsed	NA	NA
		rehearsed	151	48.00
	Act 8	reference	236	53.60
		unrehearsed	151	55.79
		rehearsed	140	54.02
Case E	Vic E	reference	NA	NA
		distress	34	8.55
	Act 09	reference	232	47.90
		unrehearsed	55	12.81
		rehearsed	59	15.59
	Act 10	reference	232	40.46
		unrehearsed	57	12.73
		rehearsed	57	13.60
Case F	Vic F	reference	NA	NA
		distress	5	4.19
	Act 11	reference	232	43.64
		unrehearsed	5	3.17
		rehearsed	5	6.79
	Act 12	reference	232	45.15
		unrehearsed	7	5.58
		rehearsed	8	6.91

5.3.1 Reference vs. distress in victims

Table 5-7 shows that for both Victims A and B, a decrease in AR is found in distress speech conditions. Both exhibit a reference AR that is within the range of typical speech rates for English speakers, which are taken to be between 4 and 6 syllables per second, see e.g. Goldman-Eisler (1968), Cruttenden (1986) and Laver (1994). Both victims exhibit a decrease in AR in distress conditions, with rates that are below the expected range. This is particularly noticeable in the case of Victim A, who shows a dramatic drop from 6.35 sylls/s in his reference material to 1.47 sylls/s in his distress material, though these measurements are based on very brief extracts. The remaining victims, however, also display distress AR values below those of expected speech rates for everyday, non-emotional speech, from 1.19 sylls/s and 3.34 sylls/s.

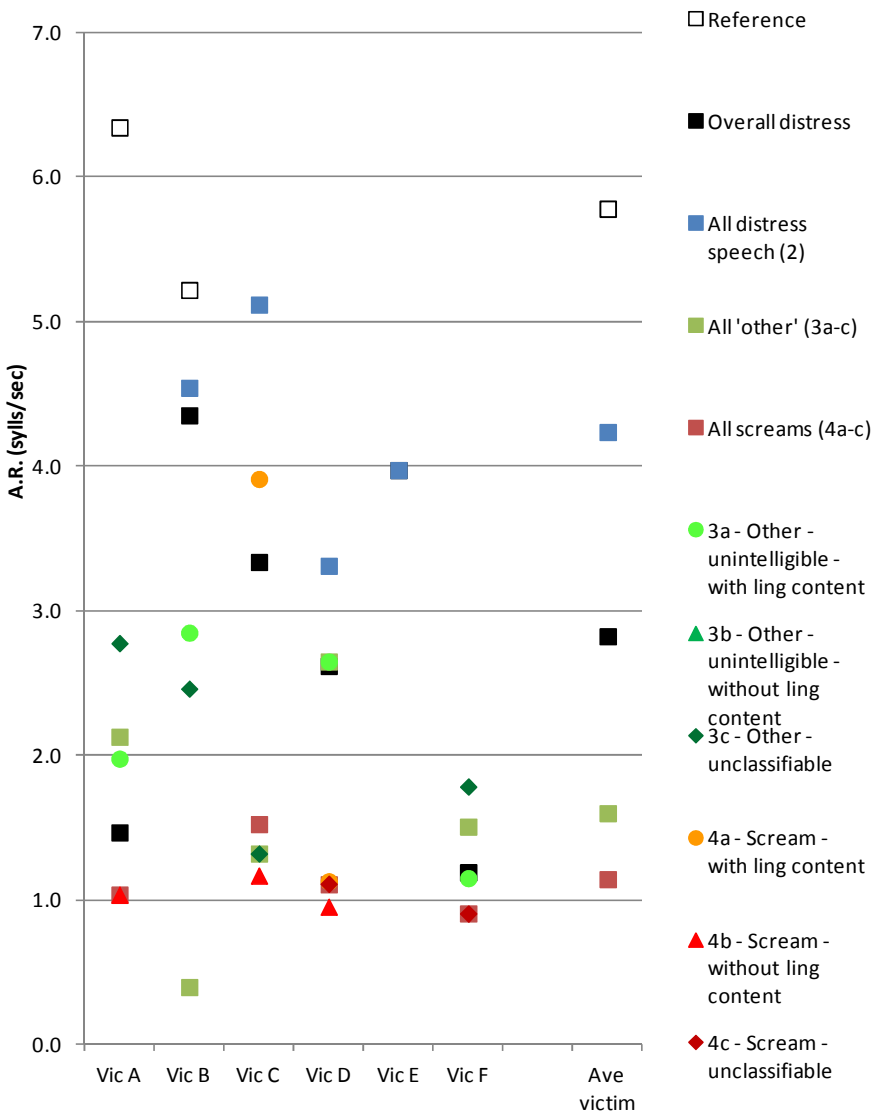
Table 5-7: Victims' AR in reference and distress conditions.

	Reference	Distress
Vic A	6.35	1.47
Vic B	5.22	4.35
Vic C		3.34
Vic D		2.62
Vic E		3.97
Vic F		1.19

Examining the victims' AR in distress productions as categorised using the taxonomy, Figure 5-35 shows that although all distress responses have a lower speech rate than that of the corresponding reference speech, there are differences between the categories. Distress productions labelled as intelligible distress speech tend to have an AR of between 4 and 5 syllables per second (mean = 4.2 sylls/s), yielding values that are within or just below the expected range of AR, whereas productions classified as 'other' or screams ranged from 0.5 to 3 syllables per second (mean = 1.6 sylls/s). Screamed productions had the lowest AR of all, with values typically around 1 syllable per second (mean = 1.15 sylls/s).

In the case of some screams and vocalisations, the AR is essentially measuring the duration of the production, not the number of screamed syllables per second. It is perhaps unhelpful to consider them in terms of AR or to compare them with the distress speech category AR These two categories markedly reduce the overall distress mean AR (i.e. the average of all distress categories). However, the use of categories does afford some descriptive acoustic information relating to screams (which are rarely defined in the literature) and in this case it can be observed that the duration of a scream is typically about 1 second. Furthermore, since the AR of intelligible speech (i.e. reference and distress speech) is often considerably higher than that of the other categories, an open question would be whether a higher AR is indicative of linguistic content (on the assumption that intelligible speech by definition contains linguistic content). Such a tendency could be relevant when assessing whether a victim had intended to produce speech with linguistic content or had unconsciously produced non-verbal vocalisations as a response to the attack. Figure 5-35 shows that Victim C has a higher AR for his scream classified as containing linguistic content (3.9 sylls/s) than for the scream without linguistic content (1.2 sylls/s). Similarly, Victim B has a higher AR for his unintelligible productions that were labelled as containing linguistic content (2.9 sylls/s) than those labelled ‘unclassifiable’ (2.5 sylls/s). However, for Victims A and F, the reverse is true, as their ‘without linguistic content’ productions (2.0 and 1.2 sylls/s, respectively) have a higher AR than those with linguistic content (2.8 and 1.8 sylls/s, respectively).

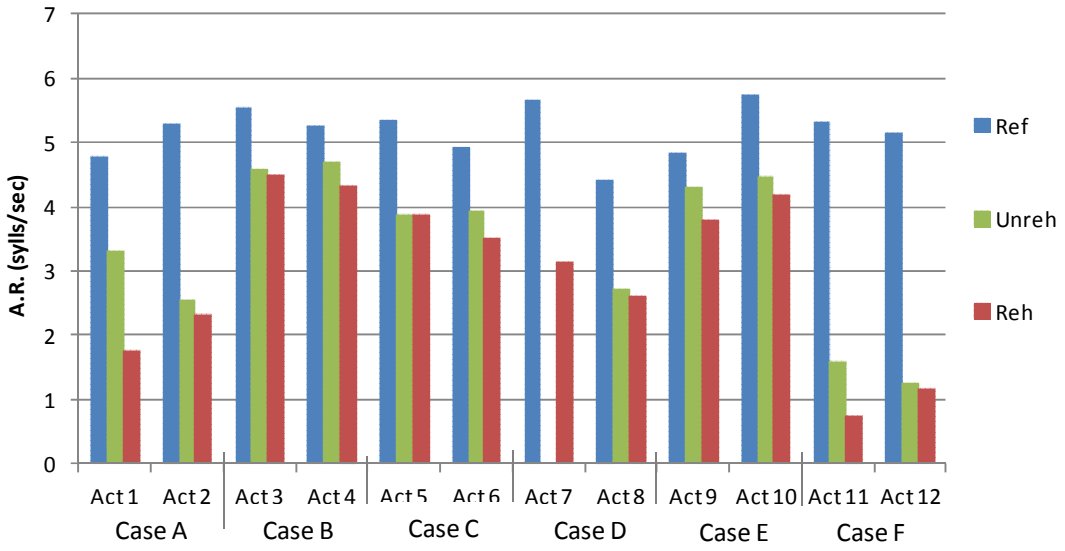
Figure 5-35: Mean articulation rates for all victims and speech categories.



5.3.2 Reference vs. distress in actors

The AR values for all actors in reference, unrehearsed and rehearsed conditions are shown in Figure 5-36. It illustrates that the reference material is in keeping with the expected range for non-emotional speech, but that rehearsed and unrehearsed distress have a lower AR For some actors, the decrease is quite small, e.g. Actor 4 has a reference AR of 5.3 sylls/s, an unrehearsed AR of 4.7 sylls/s, and a rehearsed AR of 4.3 sylls/s. For other actors the decrease is quite marked, e.g. for Actor 11, whose reference AR is 5.3 sylls/s, unrehearsed AR is 1.58 sylls/s, and rehearsed AR is less than 1 syll/s. It can be seen that for most actors, their rehearsed and unrehearsed material have similar AR values.

Figure 5-36: Articulation rates of all actors in reference, unrehearsed and rehearsed speech conditions.



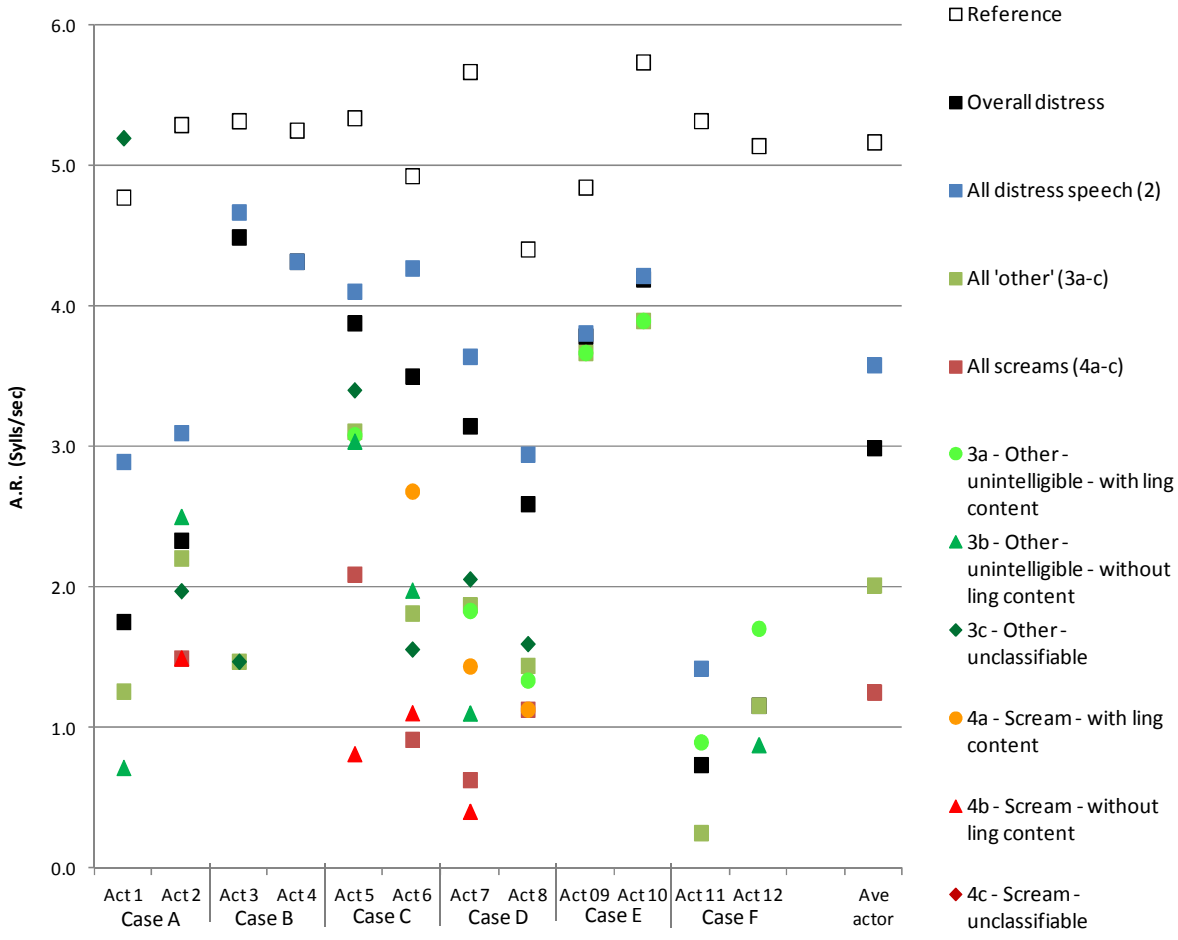
A repeated measures one-way ANOVA showed a significant effect for changes in AR across the three speech conditions (F = 27.37, df = 1.20, $p < 0.001$). As Mauchly's test of sphericity was significant, the results are reported with a Greenhouse-Geisser correction (i.e. the degrees of freedom are modified to reduce error rate). The means of AR in reference, unrehearsed and rehearsed speech conditions were 5.14 (S.D. = 0.38), 3.38 (S.D. = 1.21) and 2.98 (S.D. = 1.33), respectively. We can therefore say that actors' AR decreased from reference to unrehearsed distress speech, and then decreased further in rehearsed distress speech. Applying a Bonferroni correction (so that effects are reported at a 0.0167 level of significance), three related *t*-tests were performed, which showed that the difference between reference and unrehearsed distress speech, the difference between reference and rehearsed distress speech, and finally the difference between rehearsed and unrehearsed speech, are all significant (t = 4.94, df = 10, p = 0.001; t = 6.28, df = 11, $p < 0.001$; and t = 3.02, df = 10, p = 0.013, respectively).

Figure 5-37 compares the actors' rehearsed distress AR across the speech categories of the distress taxonomy, and shows that all distress categories have a slower AR than reference material, with intelligible distress speech having the least slow AR among the distress categories (mean = 3.6 sylls/s) and screamed productions the

lowest (mean = 1.3 sylls/sec). All unintelligible 'other' distress productions average 2 syllables per second.

A one-way repeated-measures ANOVA shows that there were significant differences between the categories (F (38.08, df = 2, $p < 0.001$)). Results from three related t-tests (performed with Bonferroni correction) reveal that differences between intelligible distress speech vs. 'other' distress, intelligible distress speech vs. screamed productions, and 'other' distress' vs. screamed productions, are all significant: $t = 4.79$, df = 9, $p = 0.001$; $t = 6.81$, df = 4, $p = 0.002$; $t = 5.34$, df = 4, $p = 0.006$, respectively.

Figure 5-37: Actors' articulation rates across speech categories in rehearsed distress.



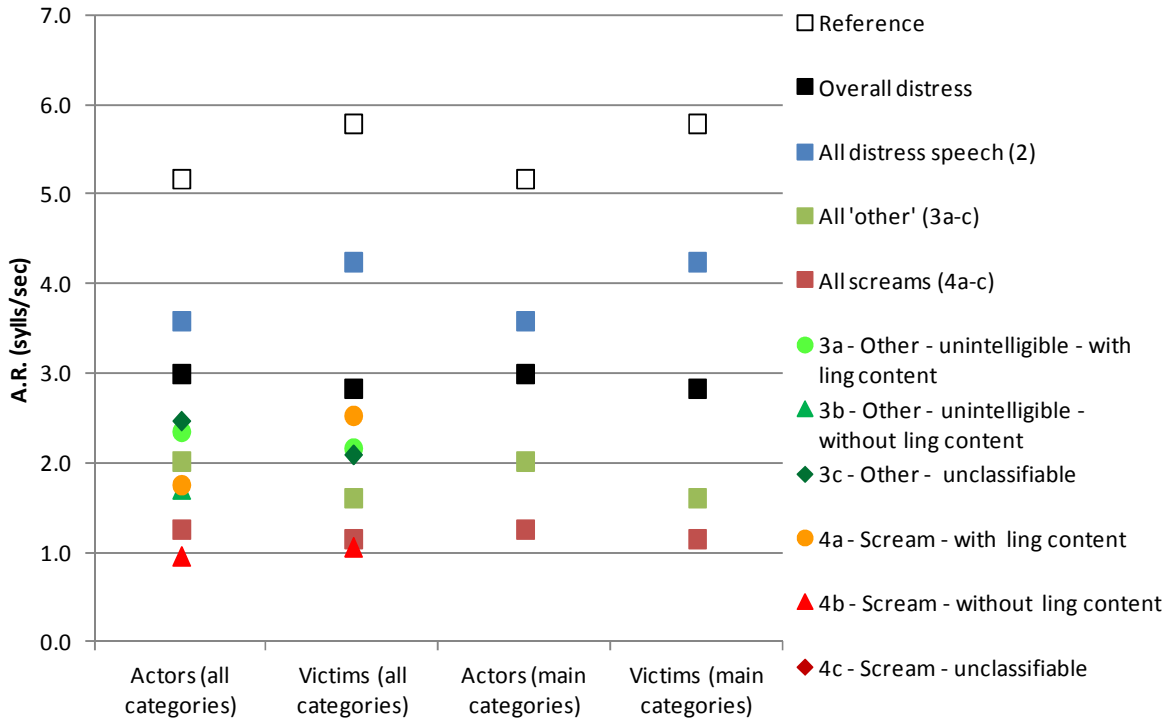
Comparing the AR of unintelligible 'other' speech and screamed productions with and without linguistic content, it can be seen that for Actors 7, 8 and 12, their unintelligible 'other' vocalisations produced with linguistic content have a higher

AR than those without. Similarly, for Actors 6 and 7, screams produced with linguistic content had a higher AR than those without linguistic content. Actors 9 and 10 had only examples of unintelligible speech classified as containing linguistic content, so no comparison can be drawn with their 'without linguistic content' productions, but the AR of the unintelligible speech with content is comparable to that of intelligible distress speech. For female actors, at least, it appears that productions containing linguistic content have a higher AR than those that were unclassifiable or labelled as having no linguistic content.

5.3.3 Actors' vs. victims' distress

Figure 5-38 compares mean AR averaged across actors and victims as well as across speech categories as defined in the taxonomy. The pattern of a decrease from reference material in all distress categories is true for both groups, and the order of the decrease across the distress categories is the same: intelligible distress speech has the least slow AR amongst the distress categories (though is typically faster than the overall, aggregate distress AR), and screamed productions the slowest. All unintelligible 'other' productions fall in between these two categories, though they are typically nearer the screamed averages than the intelligible speech ones. In both groups we see that screams with linguistic content tend to be produced at a slightly faster rate than other screamed productions; in the case of the actors their screamed productions with linguistic content fall in the range of 'other' vocalisations, but for the victims their screamed production with linguistic content has a higher AR than the 'other' productions. There was no significant difference between victims' and actors' overall distress AR, nor between their ARs calculated for each distress category.

Figure 5-38: Mean AR averaged across actors and victims and across speech categorisations.



5.3.4 Summary of AR results

§5.3 examined differences in articulation rate between reference and distress speech, and between actors' and victims' distress responses. The findings are summarised as follows:

- Both actors and victims show a decrease in AR from reference to distress material. This decrease is statistically significant for the group of actors.
- In distress speech material, both actors and victims have the slowest AR when screaming and the fastest AR in intelligible speech productions, with 'other' vocalisations having a slightly faster rate than screamed productions.
- The decreases in AR between intelligible distress speech and 'other' productions, intelligible distress speech and screams, and 'other' productions and screams, are all statistically significant for the actors.
- There was no significant difference between the actors' and victims' overall distress AR
- There was no significant difference between actors' and victims' AR for intelligible distress speech, 'other' productions and screams.

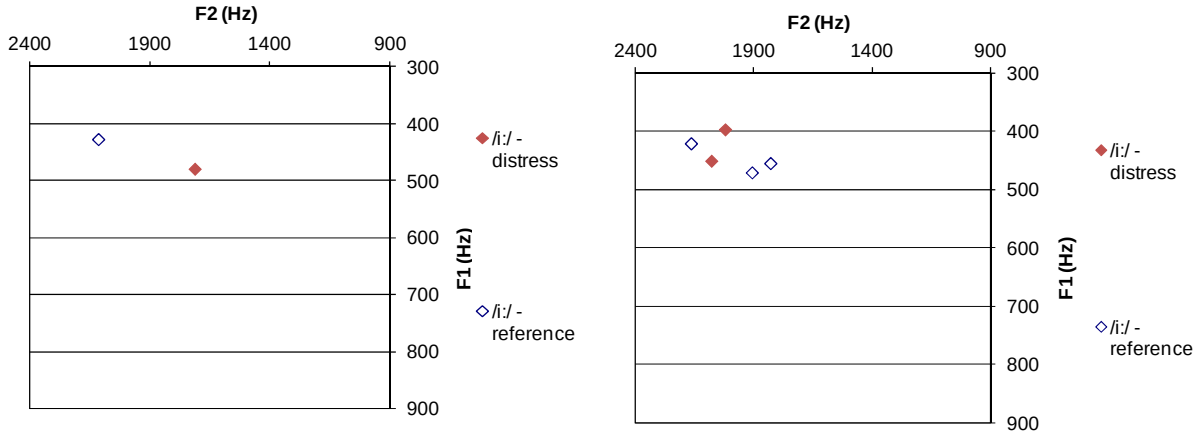
5.4 Vowel formants

Vowel formant measurements were taken for 7 monophthongal vowel categories: /iː ɪ ɛ a ɑː ɒ ʌ/. Unlike previous parameters, vowel tokens are not compared across the different categories of the distress taxonomy as there is no value in measuring vowels where speech is unintelligible or questionable. Vowel formant measurements were only analysed if the linguistic content was understandable without difficulty. Results are therefore not present across the distress taxonomy; for comparisons involving vowel formants the distinction is purely between reference and intelligible distress speech.

5.4.1 Reference vs. distress in victims

Case A has one of the shortest durations of analysable material. Consequently, there are only two vowel categories common to both Victim A and B - the monophthongal vowel /iː/ and diphthongal vowel /əʊ/. Figure 5-39 shows F1~F2 vowel plots for /iː/ as produced by both victims. For Victim A, the distress vowel is retracted (F2 decreases) and less close (F1 increases). Both vowel tokens follow a lateral consonant ('please' in distress and 'Leeds' in reference) and both are closed syllables.

Figure 5-39a(l) and 39b(r): Vowel scatter plots of /iː/ for Vic A (l) and Vic B (r)



For Victim B, the distress vowel tokens are plotted within the F1~F2 range of the reference tokens, but note that distress tokens both follow lateral consonants (the word 'bleeding' is produced twice in distress) and are in closed syllables. None of the reference tokens contains a preceding lateral, though the one reference token that has a higher F2 is a vowel token following an alveolar approximant in the word

‘three’. It is an open syllable and is more phonologically similar to the distress vowel token than the other two reference /iː/ tokens due to presence of a preceding rhotic consonant. Considering only post-liquid tokens for the comparison, a decrease of F2 (retraction) is common to both victims. The remaining reference tokens include one closed and one open syllable.

Table 5-8 shows mean formant values for vowel monophthongs from the reference and distress speech of Victims A and B. (A full list of formant values for all the victims’ vowel tokens can be found in Appendices F3 and F4). As illustrated in Figure 5-39, Victim A produces a higher F1and a lower F2 in distress speech for /iː/, whereas Victim B produces a (slightly) lower F1 and a higher F2 in distress speech for the same vowel. Victim A also produces a higher F3 in distress speech, yet Victim B produces a similar F3 in both reference and distress speech (Table 5-8). Examining the other vowel categories available in Victim B’s sample it can be observed that, on the whole, his F1 and F2 typically increase in distress speech (i.e. the vowel is typically realised more open and more front), yet his F3 remains stable across both reference and distress speech. This holds true for both front and back vowels, and vowels of differing heights, though note that there are only four vowel categories within which comparisons can be made.

Table 5-8: Mean formant values for monophthongal vowels produced by Victims A and B in reference and distress speech.

	F1		**F2**		**F3**	
Speaker	**ref**	**dis**	**ref**	**dis**	**ref**	**dis**
Vic A						
iː	428	590	2114	1712	2562	3529
Vic B	**460**	**515**	**1522**	**1590**	**2507**	**2416**
iː	449	424	1967	2050	2793	2774
ɪ	367	444	1670	1718	2465	2340
a	567	649	1345	1404	2482	2243
ʌ	455	542	1107	1189	2286	2307

5.4.2 Reference vs. distress in actors

Inspection of vowel formants comparing across reference and distress material did not reveal any salient patterns; in fact, the results seem chaotic. Table 5-9 summarises changes in vowel formant values across actors and across vowel categories (mean formant values for the actors' vowel tokens can be found in Appendices F5 and F6). It illustrates whether there is an increase, decrease or no change in vowel formants across the three speech conditions. For example, Actor 1's /iː/ vowel (fourth row) shows an increase in F1 in the unrehearsed distress material compared to his reference material (second column).

Data from Table 5-9 and Table 5-10 suggest that there is a tentative pattern in F1 which increases in unrehearsed distress compared to reference material. A Friedman ANOVA revealed significant changes amongst the actors' F1 values across the three speech conditions (χ^2 (10) = 10.4, $p < 0.03$), and Wilcoxon matched pairs tests (with a Bonferroni correction applied so that all effects are reported at a 0.0167 level of significance) found that F1 in the unrehearsed distress condition was significantly higher than the reference speech material (Wilcoxon Z = 2.432, two tailed, p = 0.01). There was no significant difference between F1 values in reference and rehearsed speech, nor in rehearsed versus unrehearsed speech. There were no significant changes to the actors' F2 or F3 frequencies across the speech conditions.

Table 5-9: Vowel formant changes across male actors for reference, rehearsed distress and unrehearsed distress speech conditions.

	F1			F2			F3		
Speaker	**ref-unreh**	**ref-reh**	**unreh-reh**	**ref-unreh**	**ref-reh**	**unreh-reh**	**ref-unreh**	**ref-reh**	**unreh-reh**
Act 1	<	<	>	<	=	>	<	>	>
iː	<	<	>	<	=	>	<	>	>
Act 2	<			<			<		
iː	<			<			<		
Act 3	<	<	=	=	>	=	=	=	=
iː	<	=	>	=	=	=	=	>	>
ɪ	>	=	<	=	=	=	=	<	<
a	<	<	<	>	>	>	=	=	=
ʌ	<	<	>	>	>	>	>	=	<
Act 4	<	=	>	<	=	>	=	=	=
iː	<	<	=	=	<	=	=	<	<
ɪ	<	<	=	<	<	=	>	=	=
a	<	=	=	<	=	>	>	>	>
ʌ	=	>	>	=	>	>	>	>	>
Act 5	<	<	=	<	<	=	=	=	=
iː	<	=	>	=	=	=	>	>	=
ɪ	<	<	=	<	<	=	=	<	<
ɛ	<	<	>	<	<	=	=	=	=
a	<	<	=	=	<	=	<	>	>
ʌ	<	<	<	=	<	<	>	=	<
Act 6	=	=	=	=	=	>	=	>	>
iː	<	<	=	=	>	>	=	>	>
ɪ	=	<	<	<	<	>	=	=	=
ɛ	=	=	=	<	=	>	=	>	>
a	<	=	=	<	=	>	=	>	>
ʌ	=	>	>	=	=	=	=	=	=

Table 5-10: Vowel formant changes across female actors for reference, rehearsed distress and unrehearsed distress speech conditions.

	F1			F2			F3		
Speaker	**ref-unreh**	**ref-reh**	**unreh-reh**	**ref-unreh**	**ref-reh**	**unreh-reh**	**ref-unreh**	**ref-reh**	**unreh-reh**
Act 7		=			=			=	
iː		<			>			>	
ɪ		=			>			<	
ɛ		<			=			=	
a		<			<			<	
ɑː		>			=			>	
ɒ		<			<			=	
ʌ		>			<			>	
Act 8	<	<	=	=	=	=	>	=	<
iː	<	<	>	<	=	>	<	<	=
ɪ	<	<	=	>	=	<	=	=	<
ɛ	<	<	>	>	=	<	>	=	<
a	>	=	<	=	=	=	>	>	<
ɑː	<	=	>	<	<	>	<	<	<
ɒ	<	<	<	>	=	<	>	<	<
ʌ	<	=	>	<	=	>	>	>	>
Act 9	<	=	=	>	>	=	=	=	=
iː	=	<	<	>	>	=	>	=	<
ɛ	<	<	<	=	=	=	=	>	>
ɒ	<	>	>	=	>	>	=	=	=
ʌ	=	>	>	>	>	=	>	=	=
Act 10	>	>	=	=	>	>	>	>	=
iː	=	>	>	>	>	=	>	>	<
ɛ	=	=	>	<	>	>	>	=	<
ɒ	>	=	<	<	<	<	>	>	<
ʌ	>	>	>	=	>	>	>	>	>
Act 11	=	>	>	<	=	=	<	<	<
iː	=	>	>	<	=	=	<	<	<
Act 12	<	<	<	=	=	=	>	<	<
iː	<	<	<	=	=	=	>	<	<

Individual Wilcoxon pair tests were conducted across speakers (Table 5-11) and vowels (Table 5-12) to further explore the findings reported in the previous paragraph. Table 5-11 shows that Actors 5 and 8 demonstrate significant changes in their distress and reference material. For Actor 5, both his F1 and F2 increased significantly in the unrehearsed distress material relative to reference speech, whereas Actor 8 showed significant increases in both her F1 and F2 in the rehearsed distress material relative to reference speech.

Table 5-11: Changes in formant values tested for significance across actors (* p < 0.05).

	F1			F2			F3		
Speaker	**ref-unreh**	**ref-reh**	**unreh-reh**	**ref-unreh**	**ref-reh**	**unreh-reh**	**ref-unreh**	**ref-reh**	**unreh-reh**
Act 1	<	<	>	<	=	>	<	>	>
Act 2	<			<			<		
Act 3	<	<	=	=	>	=	=	=	=
Act 4	<	=	>	<	=	>	=	=	=
Act 5	<*	<	=	<*	<	=	=	=	=
Act 6	=	=	=	=	=	>	=	>	>
Act 7		=			=			=	
Act 8	<	<*	=	=	=*	=	>	=	<
Act 9	<	=	=	>	>	=	=	=	=
Act 10	>	>	=	=	>	>	>	>	=
Act 11	=	>	>	<	=	=	<	<	<
Act 12	<	<	<	=	=	=	>	<	<

Table 5-12 shows that F1 and F2 typically increase in the distress speech condition (both rehearsed and unrehearsed), though /ʌ/ is an exception, whereas F3 has mixed changes across the conditions. Significant changes across vowels were few and were typically found in the front vowels. For /iː/, there were significant increases in F1 from reference to rehearsed distress, and significant increases in F2 from reference to unrehearsed distress. A significant increase in F3 from reference to rehearsed distress was found in /ɛ/, and for /a/ there was a significant increase in F1 from reference to rehearsed distress.

Table 5-12: Changes in formant values tested for significance across vowels (*p<0.05, **p<0.01).

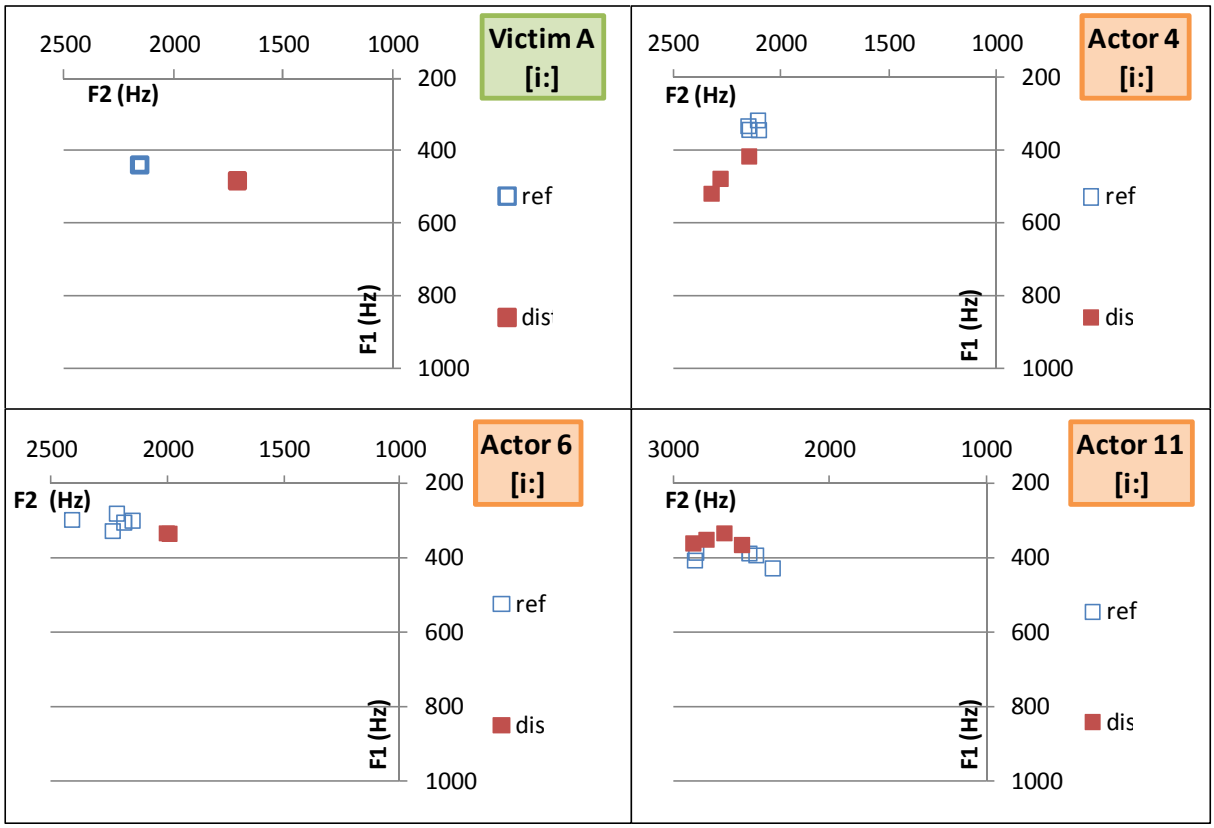
	F1			F2			F3		
Vowel	**ref-unreh**	**ref-reh**	**unreh-reh**	**ref-unreh**	**ref-reh**	**unreh-reh**	**ref-unreh**	**ref-reh**	**unreh-reh**
iː	<**	<	=	=	=*	=	=	=	=
ɪ	=	<	<	=	<	=	=	=	<
ɛ	<	<	=	=	=	=	>	=*	=
a	=	<*	<	=	=	=	>	=	=
ɑː	<	=	>	<	<	>	=	<	=
ɒ	=	<	<	=	=	=	>	=	<
ʌ	=	>	>	=	=	=	>*	>	=

Table 5-11 and Table 5-12 illustrate that some actors are more variable than others in their direction of change, if any, and occasionally these changes result in significant, if small, differences across the speech conditions. However, although some localised trends are observed, when taken as a whole no consistent general pattern emerges. Given this lack of consistency both within and across actors, it is not clear how to form a generalisation based on the current data. More data would be required to clarify this situation.

5.4.3 Distress in actors and victims

Figure 5-40, showing F1~F2 vowel plots of /iː/, the only vowel common to both victims and actors, highlights the complexity of the vowel formant findings by illustrating the variety of patterns found across the speakers; no single trend can be observed. As reported in §5.4.1, the pattern found for the two victims with reference material available for comparison was one of a retraction and centralisation of the /iː/ vowel in distress speech (characterised by an decrease in F2 and an increase in F1). This pattern is observed in some actors, e.g. Actor 6. However, tendencies in other directions are also observed across the actors. For example, Actor 4's distress /iː/ vowels are more open than those in his reference speech (F1 increases) whereas Actor 11's distress /iː/ vowels are closer (i.e. F1 decreases) and clustered closer together than in her reference speech.

Figure 5-40: Vowel plots for /iː/ demonstrating lack of uniformity in changes from reference to distress speech.



Although few formant data were available from victims, separate Mann-Whitney tests were conducted for the first three formants of /iː/ to test for differences in distress formant data across actors and victims. There were no significant differences between actors' and victims' /iː/ formant values in distress speech.

5.4.4 Summary of vowel formant results

The results of the vowel formant analysis reveal that:

- Victims A and B had a similar pattern of retraction in /iː/ in a post-liquid environment.
- Changes were observed in vowel formants across reference, rehearsed distress and unrehearsed distress conditions and across actors, but changes were not always in the same direction, and few were statistically significant.
- There was, however, a significant increase in the actors' F1 in unrehearsed distress when it was compared with their reference material.

- For /iː/ in the acted data, there were significant increases in F1 from reference to rehearsed distress, and significant increases in F2 from reference to unrehearsed distress.
- There were no significant changes between the actors' /iː/ formant data and the victims' /iː/ formant data.

5.5 Chapter summary

This chapter has illustrated that F0 mean and standard deviation increase in distress speech for both actors and victims, and that AR decreases in distress speech for both actors and victims. Intensity decreased in the distress speech produced by actors. Vowel formants also changed in the distress speech of actors and victims, though not always in the same direction or to the same extent. The relationship between acoustic and perceptual correlates of distress remains unclear. Acoustic parameters can be used to distinguish between reference and distress conditions for actors and victims, but are not so helpful in discriminating between actors and victims. Forensic practitioners should therefore continue to refrain from making psychological assessments of distress, or at least to exercise caution when doing so, if their analysis is based solely on these four acoustic parameters.

6. Methodology of Perceptual Experiment

This chapter provides a description of a perceptual experiment used to investigate perceptual cues to acted versus authentic distress. Section 6.1 introduces the research questions that the experiment seeks to address based on the findings from the acoustic study of distress productions. Section 6.2 describes the stimuli used in the perceptual experiment and section 6.3 provides information about the listeners who participated in the experiment. Section 6.4 highlights the experimental design of the perceptual test, while section 6.5 clarifies the procedure used throughout the experiment. A summary of this chapter is provided in section 6.6.

6.1 Research questions

The findings reported in chapter 5 of this thesis show that significant acoustic differences were apparent between reference and distress speech in terms of F0 (mean and range), and articulation rate, while distress productions by actors and victims could not be consistently differentiated between the two groups. From these results we could infer that acted and authentic distress may not be easily distinguishable. However, these acoustic parameters represent only a handful of potential parameters on which listeners may base an opinion of the authenticity of distress productions. After all, it is an apparent paradox that perceptual studies may reveal accurate levels of emotion identification, yet production studies often lack an identifiable set of vocal cues that reliably differentiate between emotions (Scherer 1986). Moreover, we may also ask whether audiences with varying levels of familiarity and exposure to authentic distress data may be more successful at distinguishing between acted and authentic distress than the average person. To explore these issues, and address the second principal research question (§1.2), a perceptual experiment was conducted to compare brief extracts of authentic and acted distress productions that were played to three different audiences: lay people, police call takers, and forensic practitioners.

6.2 Experiment stimuli

Brief extracts from recordings of the actors' and victims' distress productions were presented as audio stimuli for the perceptual experiment.

6.2.1 Re-recording of acted data

Since the acted data were recorded using higher-quality equipment and in more controlled conditions than the real victim data, the actors' rehearsed distress recordings were degraded by recording them through a mobile telephone. One unrehearsed recording for a control extract was degraded in the same way (see §6.2.2). This was done in order to introduce noise and random effects into the recordings, thereby bringing them into line with real forensic data that is often of poor recording quality. Degradation of the acted material was designed to make it more comparable with the authentic recordings. An additional consideration was that degradation of the acted materials would also prevent participants from basing their decisions on differences in recording method and quality, allowing them instead to focus on the main aim of the experiment, which was to distinguish between real and acted distress.

The re-recording process involved the following steps:

1. Playing the original acted distress recordings from a laptop through external PC multimedia speakers to a mobile phone

2. Recording, into Sony Sound Forge, the sound signal from the mobile phone to a digital soundcard using a landline phone.

The first stage involved selecting loud-speakers that would allow playback of the original sound files without producing Global System for Mobile telecommunications (GSM) interference. GSM interference is added to the speech signal during the recording process if the interference is present in the recording equipment (Rosengren & Nilsson 1999). The interference is produced by the phone while the phone is transmitting. It induces a varying current in other audio equipment components due to changes in the electromagnetic field, especially if the equipment and phone are in close proximity to each other (Harrison 2001). Although this type of interference is often present in forensic recordings, it is not present in the authentic extracts that are used in this perceptual experiment. It was therefore important to avoid GSM interference being present in the acted extracts. Avoiding GSM interference proved quite difficult as the loud-speaker and mobile phone needed to be close to one another in order for the sound signal to be picked up by the

mobile phone, resulting in unwanted GSM interference noise being recorded as well. After trial and error using various speaker sets available to me, a pair of Trust Soundforce 2.0 (Speaker Set SP-2200) PC multimedia speakers (single driver, wide bandwidth) were used, as they allowed for close proximity between the speaker and mobile phone without interference. One speaker was muted in order to more similarly replicate a real-life situation between a human speaker and mobile phone (where there is only one sound source in human speech), and the other was positioned approximately 1-2cms away from the mobile phone receiver using a clamp stand to mimic the position of a real human speaker using a mobile phone.

Much consideration went into which mobile phone should be used, since the original victim cases dated from the mid 1990s to mid 2000s, a period when mobile phone technology was not as developed as it is now. I wanted to ensure that the mobile phone(s) used to re-record the sound files were suitably basic so as to mimic the quality of technology available at the time. Until the recent advent of iPhone and Android mobile phone technologies, Nokia had a monopoly on mobile phone sales in the world. Their most basic model, the Nokia 1100 (released 2003) is the highest-selling mobile phone in history, but a functioning model could not be found for the re-recording of material. A functioning Nokia 2310 (released 2006) and a Motorola V500 (released 2003) were sourced for the purpose of re-recording. Both these phones could be considered typical of the early to mid 2000s, but are basic nowadays due to their lack of functions such as good quality camera, internet access or music capabilities. Two phones were used for re-recording in order to introduce additional variability amongst the re-recorded rehearsed distress files and to avoid any idiosyncratic recording features that might help listeners to distinguish the re-recorded acted material from the authentic data.

The second stage involved making sure the mobile phone was connected to a receiving telephone with its handset muted (Audioline Business Class AUB1 telephone and Prospect TC30 Telephone Interface Adapter handset with use of a Rane RS1 230 Vac remote power supply and a Rane MS-1 Microphone Preamplifier) before playing back the original soundfiles through the mobile phone. These soundfiles were then re-recorded digitally as a received phone signal directly

into Sony SoundForge. A full list of the re-recording equipment can be found in Appendix D2. The handset microphone of the Telephone Interface adaptor was muted to avoid extraneous noise being picked up at the other end of the recording process.

The re-recording of acted distress data took place in environments chosen to replicate, as far as possible, the original cases. Cases A and B involved attacks outside, and so the re-recording took place outside in a car park not far from a main road. Cases C to F involved indoor attacks and so were recorded in a living room, bathroom, hallway and bedroom respectively. Where original authentic case material included descriptions of room sizes and locations, the dimensions of each room used for re-recording were taken into account where possible. For each case there were two acted rehearsed distress sound files. One was recorded using the Motorola phone, the other the Nokia. One exception was for Case D, in which both acted re-recordings used the Nokia phone, as there was no reception using the Motorola handset in the bathroom (Appendix D2).

While I was responsible for playing back the original sound files, switching between mobile phone devices, and changing recording locations, a fellow PhD student (also trained in the analysis of forensic audio) listened to the resulting re-recorded rehearsed distress files live using headphones and was able to indicate if a file had to be re-recorded due to, for example, excess background noise impeding audibility or one-off loud noises such as sirens and door slams occurring nearby being picked up in the re-recording.

6.2.2 Selection of Stimuli

Where possible, screamed and distress speech productions were identified in the recordings of each of the six victims, and then their acted counterparts from both actors were also selected. Since there were two actors for every victim, only one actor's corresponding distress material was used per distress speech stimulus, but both actors were always represented as there were always two or more stimulus extracts per victim. The corresponding acted stimuli were chosen randomly in that for each real distress stimulus, a coin was tossed to decide between the two actors. If there were multiple stimuli per victim and the coin tosses for each of these extracts

produced the same actor, then once that actor had achieved half of all the available stimuli selected, I would then select the other actor for the remainder to ensure that both actors appeared in the experiment.

For Cases A and F, which had less material to provide for stimuli than the other cases, both actors' corresponding stimulus material was used. Likewise, where only one scream was produced by a victim, both actors' representations of it were included in the experiment in order to increase the number of responses. This meant that the experiment was weighted slightly more towards acted extracts than authentic extracts (56% versus 44%). Table 1 below shows the proportions of different types of stimuli extracts and the cases from which they originated.

Table 6-1: Proportion of different types of stimuli in experiment.

	Screams		**Distress speech**	
Case	**Real**	**Acted**	**Real**	**Acted**
A	1	2	1	2
B	0	0	3	3
C	1	2	2	2
D	2	2	2	2
E	0	0	3	3
F	1	2	1	2
Totals	*5 (13%)*	*8(20%)*	*12 (31%)*	*14 (36%)*

In addition to the 39 extracts listed in the table above, 4 further extracts were selected as controls. Ideally, these controls were to act as a benchmark by containing examples of non-controversial victim extracts and non-controversial actor extracts to check that the participants had similar perceptions of non-controversial acted and authentic distress, before examining their perceptions of disputable distress productions. However, since listeners' perceptions of sounds of real distress are potentially subjective and are indeed the subject of this investigation, no one extract could be selected as a control to make sure it could be identified unanimously as a production by a real victim. One example did exist, though, of what I and an experienced forensic phonetician considered non-controversial acting, i.e. acting that is recognisable for not being realistic, and so this was included as a control. It should

be noted that this was in no way a reflection of the skills of the actor involved; it concerned a misprint in the script in which the actor was unaware that she was meant to respond until after an awkward (unscripted) silence. The actor, in a bit of flurry, then read out the word 'argh' rather than taking it to represent a scream/sound/vocalisation of their choice (the actors were told they could choose to portray non-linguistic script material in any way they pleased and that they were able to improvise from the script should they want to). The misprint occurred in the very first run through of this script in an unrehearsed performance and was rectified for future performances. Consequently, actors were reminded that the words on the script were open to interpretation and it was up to them about how to portray 'words' such as 'argh'. Three other controls were selected, one from another actor and two from different victims. They were played at the beginning and also the middle of the experiment, and together the four were used to monitor participants' responses as they became more exposed to the data. A list of all the extracts, including the section of speech they contain as well as the case from which the section originates, is provided in Appendix G4.

6.2.3 Re-recording problems

Throughout the re-recording and stimuli selection process, it became apparent that some acted extracts contained more echo than others due to room reverberation in the original acted recording. This was not typical of all acted extracts, but it was more prevalent in acted material, especially in cases which were meant to be taking place outside. Moreover, in some of the quieter acted recordings, a soft, rhythmic, ticking noise, the source of which is unknown, was re-recorded when played loudly through the mobile phone. This was not particular to a specific model of mobile phone, only to recordings which had to have the volume increased for re-recording purposes. A further concern was that with the exception of one actor, the actors all delivered their distress material in SSBE, whereas the victims had mainly regional British accents. Before conducting the main perceptual experiment, I first played the extracts informally in random order to some test listeners - a trained phonetician, a trained socio-phonetician and a naïve listener - to see if any extracts stood out. I was deliberately vague about cues they should be listening for, but information that was to be presented to the experimental participants, such as the fact that extracts

originated from different speakers and locations, e.g. outside and indoors, and the fact that speakers would be from a variety of places across the British Isles, was also conveyed to the test listeners, and they all knew that the experiment concerned distinguishing acted from real distress. (The information sheet that was provided to participants is provided in Appendix G1). When played first in random order, the trained socio-phonetician and naïve listener had no comments to do with recording quality or regional accent; the main criticism was that some extracts sounded less 'human' (i.e. natural) than others, e.g. extract D (male) and extract AM (female). When checking the origin of these particular extracts, it was found they were from the authentic dataset and so they were left in the experiment as examples of distress produced by a victim. The trained phonetician did notice reverberation in some recordings but associated this with different data recording locations rather than to a particular dataset. However, when played grouped together by case and speaker, differences between the acted and authentic extracts, such as the soft ticking noise and the reverberation, were more noticeable to all test listeners. Regionality of accents was not commented upon.

6.3 Participants

In order to investigate whether exposure to and familiarity with authentic distress data might affect listeners' responses, three groups of participants were targeted to do the experiment - lay people, practising forensic practitioners, and police emergency call takers.

6.3.1 Lay people

Twenty lay people acted as a control group. They were recruited through the 'snowball' method (Milroy & Gordon 2003: 32) and took part in the experiment between July and November 2011. Eleven women and nine men, aged between 21 and 68 years old (mean = 38.5 years, median = 30 years) took part. With the exception of one male who was at the early stages of a linguistics PhD (without a forensic orientation), none of the lay people had had exposure to phonetics or real-life 999 calls. Some stated that they had heard real emergency calls on TV crime programmes such as *Crimewatch* (though it was pointed out that these programmes often use reconstructions of events and so the material may well have been acted)

and some had had first-hand experience of calling the emergency services, typically for witnessing car accidents and vandalism. Most participants had been educated to university level, with some having postgraduate degrees, though two females were not educated beyond 'A level'.[17] All were native speakers of British English and the majority of lay participants (13) were Northern English speakers from either Yorkshire or County Durham. The remaining seven came from other parts of Great Britain.

6.3.2 Police call takers

Fourteen police call takers were recruited through a contact in a police call centre based in a nearby county. I travelled to the Force Communication and Control Centre (FCCC) call centre and was able to sit in on a couple of shifts so that when lines were quiet, members of the police call centre team could volunteer to do the listening experiment. At the FCCC, call takers receive both emergency and non-emergency calls, though only one or two call takers are designated emergency call takers at any one time. They had all had experience with both emergency and non-emergency calls. Police call takers were aged between 23 and 61 years (mean = 41.5 years, median = 43 years) and nine of the fourteen were female. They were all native speakers of British English and nine of the call takers were from the local county.

Contacts in the fire and ambulance emergency services were not sought (even though they too are often a first point of contact to distress situations) as police are often considered the default emergency service. If the caller and his/her emergency are not immediately understood, police will be sent ahead of other emergency services. Furthermore, if an emergency situation requires more than one emergency service, the police will also be sent to the scene regardless of whether they suspect criminal involvement. Police officers and call takers are equally familiar with medical and fire emergencies as they are with their own work.

[17] The Advanced Level (A Level) is a secondary school leaving qualification in the UK which is typically taken at the end of the academic year in which the student turns 18-years-old.

6.3.3 Forensic practitioners

Twelve practising forensic experts participated in the experiment. The experiment took place at the 20th Annual Conference of the International Association of Forensic Phonetics and Acoustics in July 2011 in Vienna and the 17th International Congress of Phonetic Sciences in August 2011 in Hong Kong. A larger group was invited to take part prior to these events, though some declined the invitation. Given the specialist field of forensic phonetics, I took the opportunity to run the experiment at these conferences since the relevant pool of potential participants would be in the same place at the same time. Ideally, this group of listeners would have been all speakers of British English to match the lay people and police call taker groups. However, given that it is a small specialist field, I would have had too few participants for this group and therefore the criteria were changed to include speakers of other native languages/English dialects who have experience of teaching phonetics in Britain and/or practical casework experience involving British English speakers. The majority of forensic phonetician participants were male (9 of 12), and were aged between 27 and 64 years (mean = 45 years, median = 44years). Practising forensic experts who had taken part in the experiment described in chapter 4 were excluded from participating in this experiment due to their prior exposure and familiarity with the authentic material.

6.4 Research Design

The main goal of the experiment was to ascertain listeners' ability to distinguish between real and acted distress. Rather than choosing non-corresponding extracts from actors' and victims' recordings (where semantic content might influence listeners' responses), corresponding material for extracts from both datasets was chosen.

The experimental design uses a system in which each extract is presented randomly and as an individual extract, repeated once (i.e. played twice), for listeners to respond to and answer any questions on the response sheet provided, before moving to the next extract. A closed-set experimental design was considered, in which all versions of the same extract could be played to the listener before the listener made a decision about, for example, which extract was most likely to have been produced by

a real victim. This would have made sense in view of the fact that all authentic extracts had either one or two acted counterparts (depending on whether the test extract was compared to just one or both actors). However, given the acted data re-recording concerns expressed in §6.2.3, a random order experiment in which judgments are made based on individual extracts rather than groups of similar extracts was thought to be more appropriate. It avoided drawing attention to differences between speakers' recording environments and/or sociolinguistic characteristics, and instead encouraged the listener to assess the authenticity of the extract.

Two blocks were created from the 39 non-control extracts. One block, "Pool α", contained 20 extracts, and the other, "Pool β", contained 19 extracts. Each pool contained a mixture of acted and authentic stimuli. For the 12 authentic extracts which had only one acted counterpart, one extract would be in Pool α, and its counterpart would be in Pool β, with an equal mix of acted and authentic stimuli extracts appearing in each pool. For the 5 authentic extracts which had two acted counterparts, both acted counterparts would appear in the same pool. Two of the 5 extracts appeared in Pool α (with the 4 corresponding acted extracts presented in Pool β) and the other 3 in Pool β (with the 6 corresponding acted extracts in Pool α). Both pools featured in the experiment so that participants were able to listen to all extracts. As illustrated in Table 6-2 and Table 6-3 below, two experiments were run, varying the order of presentation to control for a possible order effect. Listeners therefore heard extracts from both pools, but listeners taking part in Experiment A heard extracts from Pool α before hearing those from Pool β. Listeners taking part in Experiment B heard extracts from Pool β before they heard the extracts from Pool α.

The blocks were designed to allow investigation of the effect on participants' decisions of order of hearing an authentic or acted extract first. The pools were organised semi-randomly in that extract numbers were generated randomly using an online randomiser, but if the subsequent extract belonged to the same case as the previous extract, it was moved to the end of the pool so as to avoid either similar words or the same voice appearing twice in succession. The 4 control extracts were

played at the very beginning of the experiment, and repeated in between the two pools.

Table 6-2: Experiment A design

Block A	Block B	Block C	Block D
Controls: 4 extracts	Pool α: 6 victim extracts (6 x 1 acted counterpart in Pool β) 6 actor extracts (6 x 1 victim counterpart in Pool β) 2 victim extracts (2 x 2 acted counterparts in Pool β) 6 actor extracts (3 x 1 victim counterpart in Pool β)	Controls: 4 extracts	Pool β: 6 victim extracts (6 x 1 counterpart in Pool α) 6 actor extracts (6 x 1 counterpart in Pool α) 3 victim extracts (3 x 2 acted counterparts in Pool α) 4 actor extracts (2 x 1 victim counterpart in Pool α)

Table 6-3: Experiment B design

Block A	Block B	Block C	Block D
Controls: 4 extracts	Pool β: 6 victim extracts (6 x 1 counterpart in Pool α) 6 actor extracts (6 x 1 counterpart in Pool α) 3 victim extracts (3 x 2 acted counterparts in Pool α) 4 actor extracts (2 x 1 victim counterparts in Pool α)	Controls: 4 extracts	Pool α : 6 victim extracts (6 x 1 acted counterpart in Pool β) 6 actor extracts (6 x1 victim counterpart in Pool β) 2 victim extracts (2 x 2 acted counterparts in Pool β) 6 actor extracts (3 x 1 victim counterpart in Pool β)

6.5 Procedure

Participants were asked to read an information sheet (Appendix G1) prior to taking part in the experiment. In most cases this was emailed to the participant in advance of the experiment. They were invited to raise any questions or concerns before agreeing to take part and signing the consent sheet (Appendix G2). They were also informed that they could opt out of doing the experiment at any time without giving

a reason. Instructions and information were also given verbally before the experiment began.

The experiment was delivered via PowerPoint and participants were provided with closed cup headphones (typically Sennheiser HD 280 Pro headphones) in order to listen to the audio stimuli. The experiment was run in PowerPoint in preference to other experimental software, e.g. Praat's Experiment Multiple Forced Choice (MFC) software, as it is the least specialised and probably most familiar of potential programs, and at the time of creating the experiment it was unclear what computer facilities and software would be available at the locations where the experiment took place.

A response sheet was also provided. This listed 4 principal questions per extract (Appendix G4). Participants were requested to:

1) specify whether they perceived the extract to have been produced by a victim or an actor using a 5-point (non-numerical) scale;

2) rate their level of confidence about their previous assessment ratings using a 5-point (non-numerical) scale;

3) state whether they perceived the extract to have been produced by a male or a female speaker;

4) provide a note if the extract was perceived to be unusual or had any noticeable features that might have influenced the participant's decision.

The phrasing of the first question on the response sheet deliberately omitted the words 'authentic' and 'genuine' as they were considered too vague and subjective, and therefore likely to vary in terms of interpretation amongst individuals. Instead, the question was presented as a choice between a real victim and an actor, where the listener could choose between 'definitely victim', 'probably victim', 'no decision', 'probably actor' and 'definitely actor'. A 5-point scale was selected in order to avoid forcing the participants to make a choice. The 'no decision' option was included

since forensic practitioners have the option of rejecting forensic material deemed unsuitable for analysis. Extracts which lead to ‘no decision’ responses may be just as informative as those that receive “definite” responses.

Participants were able to advance through the experiment at their own speed but were not allowed to repeat sound files nor to return to previous extracts to change responses.

6.6 Chapter summary

This chapter introduced a perceptual experiment designed to investigate different audiences’ perceptions of authentic and acted distress. The audio stimuli were based on the data sets presented in chapter 3, and a description was given of how the stimuli were selected. The selection and recruitment of participants were also described. Finally, the design and procedure of the experiment were explained. The findings of the experiment are presented in the next chapter.

7. Findings of Perceptual Experiment

This chapter presents the results of the perceptual experiment described in the previous chapter. It is divided into five sections. The first section examines the listeners' ability to distinguish between authentic and acted distress, and also whether familiarity with forensic material improves listeners' accuracy. The second section reports on the listeners' level of confidence when assessing authentic and acted distress, and investigates whether the listeners with more experience with forensic data are more or less confident than other listeners. The third section describes listeners' ability to differentiate between male and female voices in distress. The fourth section contains observations concerning listeners' verbal responses about why they judged samples the way they did. A summary of all the perceptual results is provided in the final section.

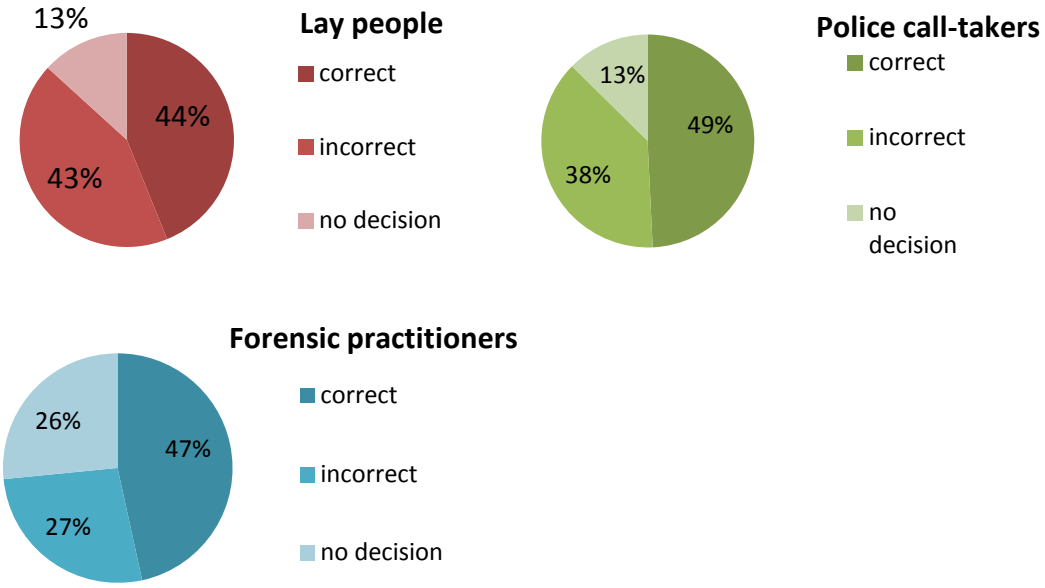
7.1 Distinguishing between authentic and acted distress

The first half of this section focuses on findings based on the listeners' responses to the experiment extracts (a by-subjects analysis, §7.1.1 - §7.1.5) and the second half of this section focuses on the extracts themselves (a by-items analysis, §7.1.17.1.6 - §7.1.11).

The first question in the perceptual experiment asked that listeners assess the brief audio extract and categorise it as one of 'definitely victim', 'probably victim', 'no decision', 'probably actor' or 'definitely actor'. The pie charts in Figure 7-1 provide a breakdown of the listeners' responses according to participant group, taking the responses 'definitely victim' and 'probably victim' to be a correct attribution if the extract was produced by a victim, and 'definitely actor' or 'probably actor' as a correct attribution if the extract had been produced by an actor. They show that the police call takers are the best performers given their higher percentage of correct responses (49% versus 44% and 47% for lay people and forensic practitioners respectively). They also illustrate that forensic practitioners have the lowest mean rate of incorrect actor/victim attributions (27%) and the highest rate of 'no decision' responses (26%). The lay people and police call takers have the same level of 'no decision' responses (13%) but vary in their incorrect attributions, with lay people

having the highest incorrect attribution rate at 43%, whereas police call takers have an incorrect rate of 38%.

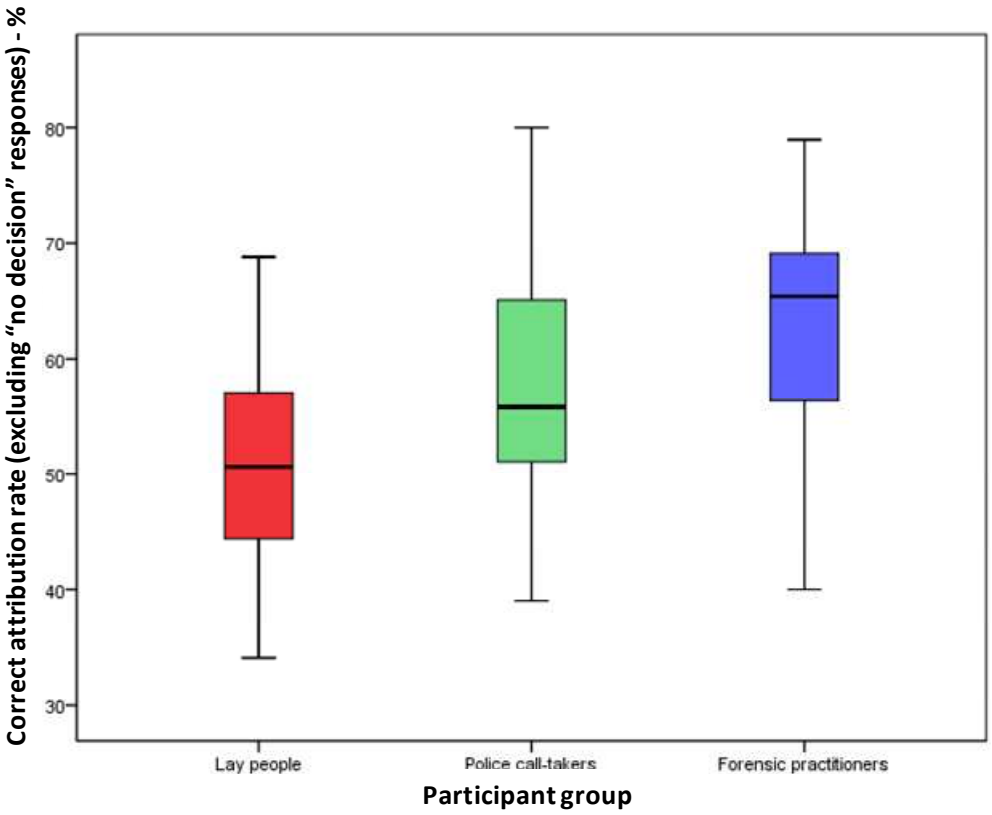
Figure 7-1: Mean breakdown of responses across participant groups.



If we take chance to be 50%, Figure 7-1 highlights that all groups perform worse than or near chance level in attributing the authenticity of the extract correctly. However, we can question whether a 'no decision' response should qualify as an incorrect response. After all, forensic practitioners have the right to refuse to analyse material if they think it is unsuitable for analysis. In some cases, exercising a 'no decision' response may be the best course of action, particularly in forensic speech science where a limited analysis may provide support for a wrongful conviction. Excluding all responses where no judgment was given, Figure 7-2 re-analyses correct and incorrect responses and shows that two listener groups, i.e. forensic practitioners (63% correct) and police call takers (57%), do perform slightly better than chance, while lay people are performing practically at chance level (51%). Figure 7-2 shows the range of correct attribution scores across the participant groups and it can be noted that the trained participants (police call takers and forensic practitioners) have similar minimum and maximum individual scores, but the forensic practitioners have a higher median (65% versus 56%). Lay people have the

lowest correct attributions in terms of individual maximum, minimum and median scores.

Figure 7-2: Correct responses (excluding 'no decision' responses) across participant groups.



The following subsections examine these findings with reference to listeners' accuracy in terms of correct, incorrect and 'no decision' responses as a function of familiarity with forensic material.

7.1.1 Correct responses

Figure 7-1 shows that the police call takers perform the best out of the three participant groups, yet the scores for each group are all very similar. In fact, no statistically significant result was found between the three groups when comparing these correct attribution scores.

However, if we exclude 'no decision' responses (as in Figure 7-2), an independent one-way ANOVA showed that there was a significant effect for participant group on correct attribution scores ($F(2, 43) = 4.77$, $p < 0.02$, $\omega = 0.43$). A linear trend, suggesting that the level of correct attributions proportionately increased from lay

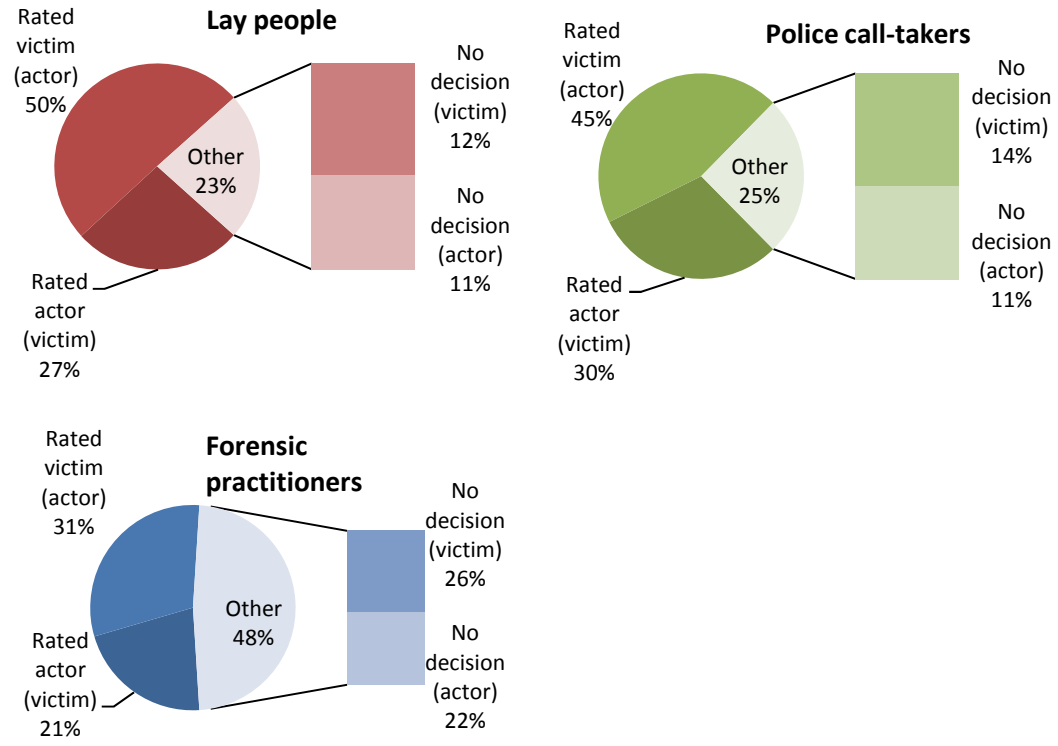
people to police call takers to forensic practitioners, was found to be significant ($F(1, 43) = 9.41$, $p < 0.01$, $\omega = 0.42$). Planned comparisons showed that trained professional participants (the police call takers and forensic practitioners) achieved significantly higher correct attribution scores than untrained listeners (lay people) ($t(43) = 2.78$, $p < 0.005$ (one tailed), $r = 0.39$) but there was no comparable significant increase between police-call takers and forensic practitioners.

7.1.2 Incorrect responses

For incorrect responses (excluding 'no decision' responses), an independent one-way ANOVA with planned contrasts again showed a significant effect of participant group ($F(2, 43) = 9.25$, $p < 0.001$, $\omega = 0.51$). A significant linear trend similar to that found for the correct response, showed that the level of incorrect attributions decreased proportionately from lay people to police call takers to forensic practitioners ($F(1, 43) = 18.34$, $p < 0.001$, $\omega = 0.53$). Planned contrasts revealed that trained professional participants, i.e. police call takers and forensic practitioners, had significantly lower levels of incorrect attributions as compared to lay people ($t(43) = -3.39$, $p < 0.001$ (one tailed), $r = 0.46$), and that forensic practitioners had a significantly lower level of incorrect attributions than police call takers ($t(43) = -2.814$, $p < 0.01$ (one tailed), $r = 0.39$).

If we now consider 'no decision' responses also to be a form of incorrect response then the picture becomes more complicated. The incorrect responses can be divided into four main types of misattribution: responses rated as 'actor' but produced by victims ('rated actor (victim)'), responses rated as 'victim' but produced by actors ('rated victim (actor)'), 'no decision' responses produced by actors ('no decision (actor)'), and 'no decision' responses produced by victims ('no decision (victim)'). Figure 7-3 provides a breakdown of all types of incorrect response and shows in parentheses the correct response (i.e. whether the sample was in fact produced by an actor or a victim). It illustrates that all participants had a tendency to mistake actors as victims rather than mistake victims as actors, and that 'no decision' responses appear almost evenly split between actor and victim mistakes. The figure further highlights the forensic practitioners' higher frequency of 'no decision' responses.

Figure 7-3: Pie charts showing breakdown of incorrect responses across participant groups.



A five-way mixed design ANOVA showed that there was a significant main effect of the type of misattribution when participants gave an incorrect response ($F(3, 129) = 48.86$, $p < 0.001$, $r = 0.52$). Contrasts revealed that misattributions where the extract was judged to have been produced by a victim but was in fact produced by an actor were more frequent than those where actors were misidentified as victims ($F(1,43) = 51.16$, $p < 0.001$, $r = 0.74$). Contrasts also revealed that misidentifications where extracts produced by actors but judged as victims were more frequent than responses produced by victims and marked 'no decision' ($F(1,43) = 85.15$, $p < 0.001$, $r = 0.82$). There was no significant difference between 'no decision' responses for extracts produced by actors and those produced by victims.

When considering 'no decision' responses as misattributions, no significant main effect of participant group was found, suggesting that the frequency and magnitude of misattribution responses as a whole were similar across all participants ($F(2,43) = 2.38$, $p < 0.001$, ns, $r = 0.29$. However, there was a significant interaction effect between the nature of misattribution and the participant group, indicating that the type of misattribution differed across participant groups ($F(6, 129) = 7.4$, $p < 0.001$, $r = 0.23$). ANOVA with planned contrasts revealed significant interactions across

participant groups when comparing misattributions where the extract was rated 'victim' but was produced by an actor, and extracts rated 'actor' but produced by a victim ($F(2, 43) = 3.96$, $p = 0.26$, $r = 0.28$). In other words, although all groups are inclined to rate actors as victims more than any other type of misattribution, the error rate is more apparent a,ong lay people. Misattributions in the other direction, i.e. victims rated as actors, are more frequent among police call takers. Misattributions in either direction are least frequent in forensic practitioners.

Another significant interaction among participant groups was found when comparing misattributions where the extract was rated 'victim' but was produced by an actor, and extracts rated as 'no decision' but produced by a victim ($F(2, 43) = 13.31$, $p < 0.001$, $r = 0.49$). It shows that responses rated as 'victim' but produced by an actor are more frequent among lay people and police call takers, and least frequent among forensic practitioners. For 'no decision' responses that were judged as having been produced by a victim, this trend is reversed; forensic practitioners are the group most frequently making this type of response, whereas the other two groups less frequently respond this way.

There was no significant interaction across the participant groups and extracts rated 'no decision', irrespective of whether they were originally produced by an actor or a victim. For both types of 'no decision' response the forensic practitioners were the group most likely to give this type of response.

7.1.3 'No decision' responses

The percentage of 'no decision' responses varied across groups in a statistically significant way ($H(2) = 12.99$, $p < 0.001$). Mann-Whitney tests were used to follow up this finding and a Bonferroni correction applied so that all effects are reported at a 0.0167 level of significance. It appears that 'no decision' rates were approximately the same between lay people and police call takers ($U = 132.5$, $r = -0.04$). However, 'no decision' responses were significantly more frequent among forensic practitioners as compared to lay people ($U = 40.0$, $r = -0.55$) and police call takers ($U = 21.0$, $r = -0.64$). We can conclude that forensic practitioners return significantly greater rates of 'no decision' responses than other participant groups. Such a

discrepancy might arise because of the fact that in forensic casework, it is in the forensic practitioner's best interests to conduct analyses only where s/he is confident that the data provide sufficient material for analysis. Otherwise, s/he may risk reaching erroneous conclusions which, if presented in court, might result in a miscarriage of justice. As part of the forensic practitioner's day-to-day job, s/he is likely to refuse to analyse cases in which there is insufficient and/or poor quality speech material, whereas the police call takers are obliged to respond to all 999 data, and lay people rarely, if ever, encounter this situation.

7.1.4 Individual performances

Figure 7-4 presents the participants' correct, incorrect and 'no decision' scores. It is arranged by highest correct scores. Participants 1-12 are members of the forensic practitioner group; Participants 101-120 are members of the lay people group; and Participants 201-214 are members of the police call taker group. Performances varied considerably across participants. It can be seen that Participant 209, a police call taker, returns the highest number of correct responses (68%), whereas Participant 6, a forensic practitioner, provides the lowest number of correct responses (26%). Among the top eight highest correct response-givers (participants achieving a correct response rate at or above 60%), 3 participants are forensic practitioners, 3 are police call takers, and 2 are lay people. 17 participants scored 40% or lower for correct responses, of which 4 were forensic practitioners, 4 were police call takers, and 9 were lay people. The overall pattern to emerge from this figure is that no single participant group dominates in terms of the highest or lowest number of correct responses. Instead, individual participants appear scattered throughout the column chart, suggesting that membership of a particular listener group does not enhance or downgrade the individual's ability to attribute the extract correctly. However, when presented by descending incorrect score, a different picture emerges (Figure 7-5).

Among the top 12 lowest incorrect scorers (those scoring less than 30% incorrect), 7 are forensic practitioners, 3 are police call takers and 2 are lay people. At the bottom end of the scale, 9 participants have 50% or more incorrect responses, of which 7 were lay people and 2 were police call takers. The forensic practitioner participants

typically have the lowest incorrect rate, whereas lay participants tend to exhibit the highest incorrect rate. The participant with the lowest incorrect response rate (12%), also has the highest response rate of 'no decisions'. The next lowest incorrect scorer is Participant 209, whose highest correct response rate, coupled with low rates for both incorrect and 'no decision' responses, mean that this participant has the best overall performance. It should be noted that no single participant stands out either positively or negative in his/her performance. While there is considerable variation, as Figure 7-4 and Figure 7-5 show, it is of a gradient nature with no outliers.

Although Figure 7-4 and Figure 7-5 confirm a few of the participant group results that were reported in the previous three sections, e.g. forensic practitioners frequently scoring amongst the lowest incorrect responses, several of the results are not confirmed, casting doubt on the strength of the participant group findings. To check whether grouping the participants according to experience is well-motivated, a cluster analysis was performed, which utilised information not just based on the rates of correct/incorrect/ 'no decision' responses (Question 1 of the experiment), but also their scores concerning confidence level (Question 2) and correct sex attribution (Question 3).

The cluster analysis was run on all 46 participants, using the variables 'no decision' responses, 'correct attribution' (excluding 'no decision' responses), 'correct sex attribution', and 'confidence level'. A hierarchical cluster analysis using Ward's method produced two clusters, which appear to coincide with familiarity with authentic data. Figure 7-6 offers a graphical illustration of the cluster analysis using the four variables described above. The coloured dashes each represent a participant: red denotes a lay person, green denotes a police call taker, and blue denotes a forensic practitioner. The first cluster was mainly characterised by a moderate correct attribution rate and a low confidence level. This is circled in Figure 7-6 and appears to contain points mainly representing police call takers and forensic practitioners (the green and blue dashes respectively). The group does not exclusively contain those with familiarity with distress data - red dashes are also visible as part of the groupings - but the majority of group members appear to be forensic practitioners or police call takers.

The second cluster is mainly characterised by lower correct attribution scores but higher level of confidences. It contains primarily lay people, but, like the previous example, does not solely contain lay people. They do, however, appear to form the majority of group membership.

It seems that the cluster analysis shows that the separating the groups by experience level is well motivated (even if not entirely predictive), as the lay people form a distinct group. The distinction between police and forensic practitioners does not emerge from the cluster analysis, perhaps indicating that the specific type of exposure to distress data that an experienced person has is less important. However, it may be that the current data set is simply not sufficiently large to separate the two.

Figure 7-4: Accuracy rates for individual listeners ordered by % correct in descending order.

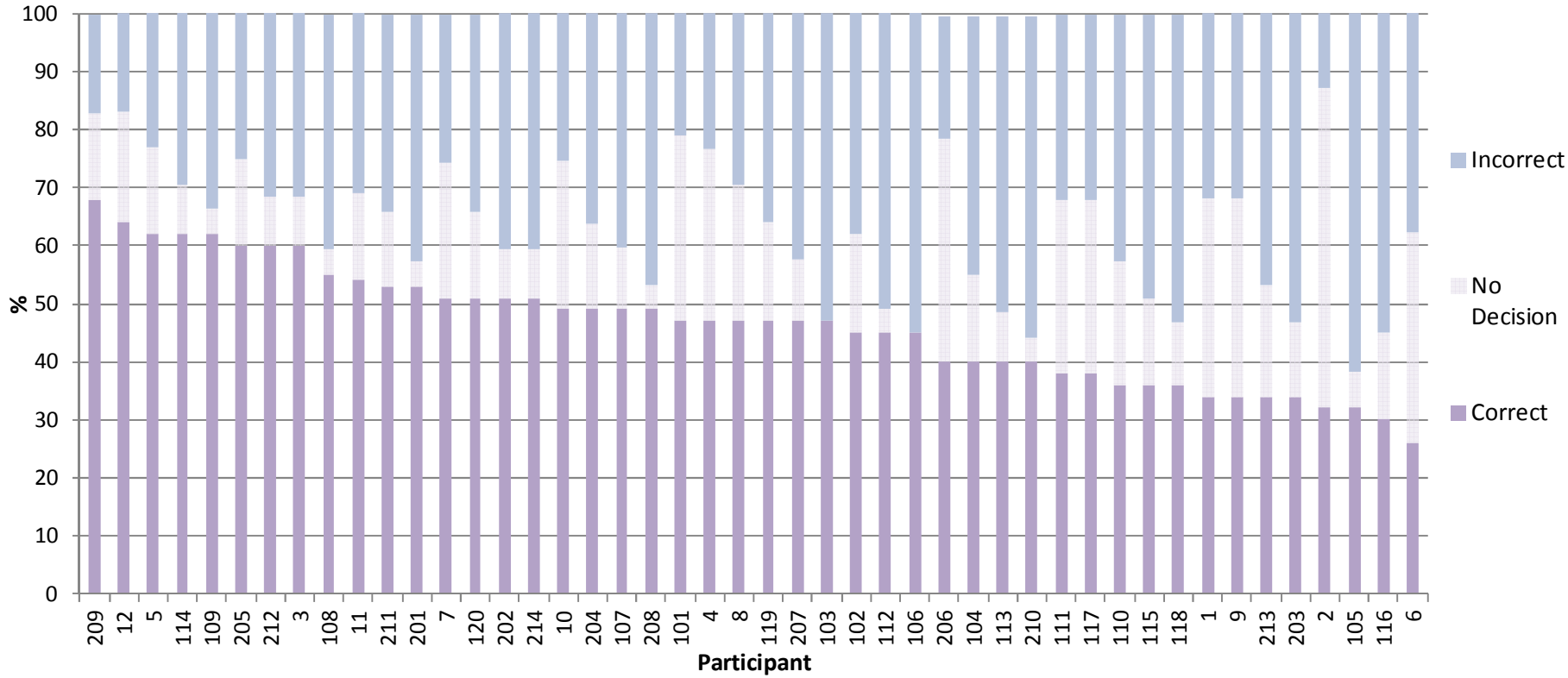


Figure 7-5: Accuracy rates for individual listeners ordered by % incorrect in ascending order.

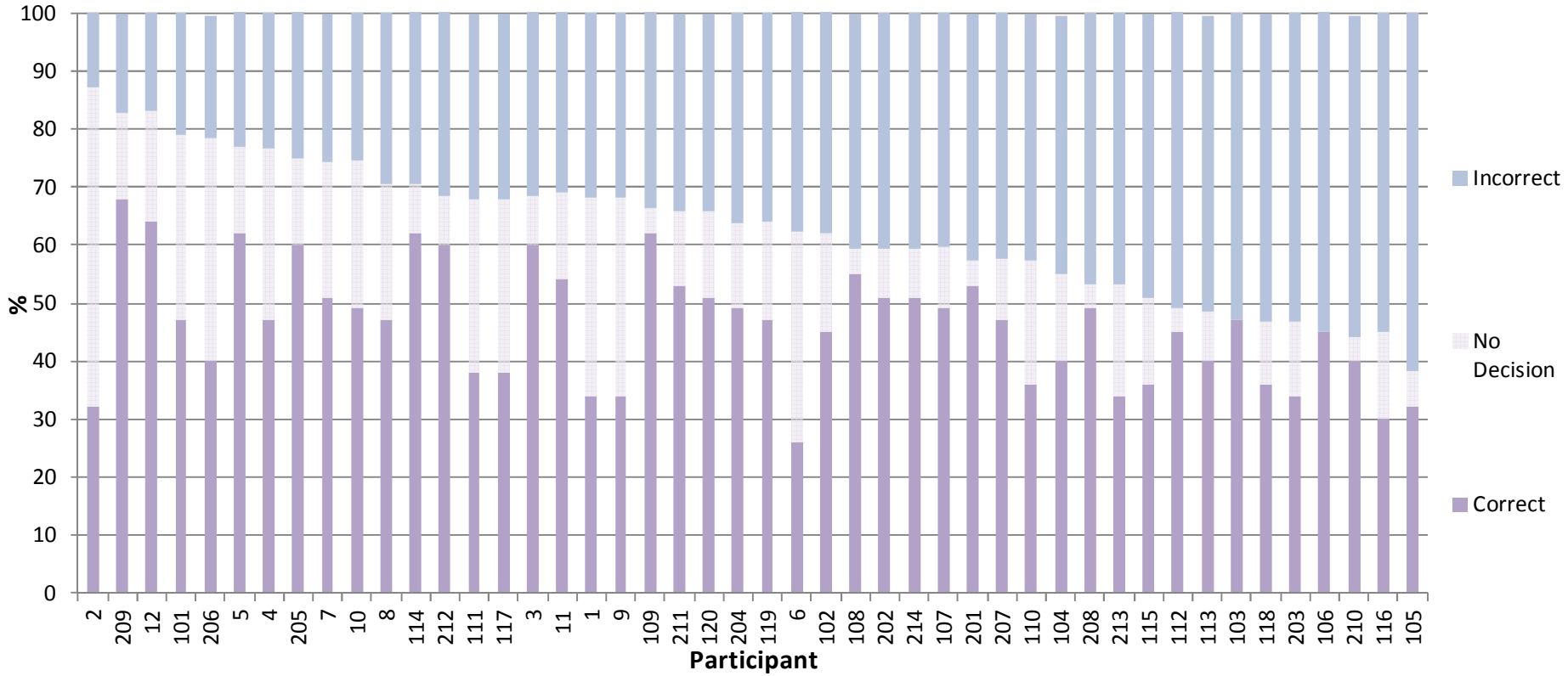
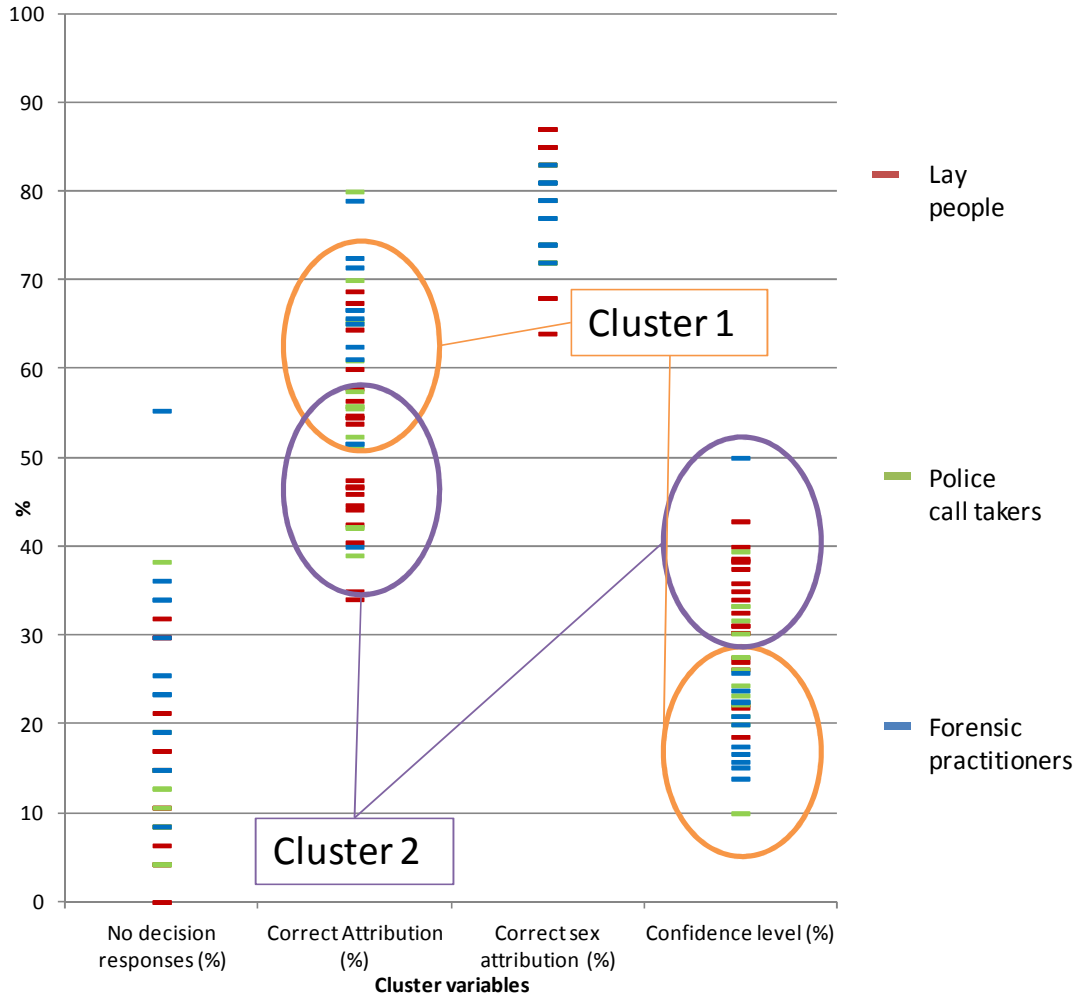


Figure 7-6: Graphical illustration of cluster groups.



7.1.5 External factors

An independent factorial ANOVA revealed no significant main effect nor an interaction between the two experiments (A or B) and the age and sex of the participant across all three participant groups. The participant's level of education was examined separately (excluding the group of forensic practitioners, as they all had postgraduate-level qualifications). A one-way independent ANOVA showed no statistically significant effect for level of education between lay people and police call takers.

7.1.6 Coding perceptual data for the by-items analysis

After having examined the perceptual experiment responses in terms of the participants' individual and group performances, the following section continues by

investigating responses to individual extracts (a by-items analysis). For each response the participant was asked to categorise each extract as one of *definitely victim; probably victim*; *no decision*; *probably actor*; and *definitely actor*. Each of the responses was coded as 1-5 for the analysis (Table 7-1). Only one response was excluded from analysis as a participant from the lay people group failed to select one of the five options for one extract (presumably in error). If the extract was produced by an actor and was marked as 'probably actor' or 'definitely actor' (a 4 or 5 when coding), it was considered a correct response.

Table 7-1: Coding for the listeners' responses used in the statistical analysis.

Options on response sheet	Definitely victim	Probably victim	No decision	Probably actor	Definitely actor
Coding for statistical analysis	1	2	3	4	5

Although no extract was unanimously identified correctly or incorrectly by all participants, there were examples of some extracts which yielded unanimous responses from a specific participant group. Moreover, some extracts showed consistent, though not unanimous, levels of correct attribution irrespective of participant group. Table 7-2 presents the extracts with consistently high and low levels of correct attribution.

Table 7-2: Acted and authentic extracts with consistent scores. (Figures in parentheses represent the percentage of correct responses for that extract; LPs = lay people, PCs = police call takers, FPs = forensic practitioners).

	Consistently Correct	**Consistently Incorrect**
Actor extract	● Extract AA (female actor, Case D) LPs (100%) PCs (86%) FPs (92%) ● Extract AP (female actor, Case E) LPs (60%) PCs (89%) FPs (75%)	● Extract AI (female actor, Case E) LPs (0%) PCs (7%) FPs (8%) ● Extract AL (female actor, Case F) LPs (15%) PCs (0%) FPs (25%) ● Extract AC (female actor, Case D) LPs (25%) PCs (7%) FPs (25%)
Victim extract	● Extract K (male victim, Case B) LPs (95%) PCs (79%) FPs (68%) ● Extract I (male victim, Case B) LPs (80%) PCs (86%) FPs (87%)	● Extract D (male victim, Case A) LPs (20%) PCs (14%) FPs (25%) ● Extract AM (female victim, Case F) LPs (25%) PCs (14%) FPs (50%)

7.1.7 Consistently correct acted extracts

Extract AA, part of a recording of a female actor (Actor 7) producing speech material from Case D, was correctly identified as an actor in 94% of all responses. The lay people were unanimous in rating this extract as having been produced by an actor, whereas the police correctly attributed it to an actor in 86% of responses and the forensic practitioners 92%. Similarly, Extract AP, a control sample from the recording of a female actor (Actor 12) producing speech and vocalisations with no linguistic content (originally scripted as a scream but misunderstood by the actor due to a typographical error in the script), scores highly as it was been consistently identified as having been produced by an actor; 73% of all responses to this extract were rated as such. The police call takers achieved the highest rate of correct

identification for this extract, at 89%, whereas the forensic practitioners attained a correct response rate of 75%, and the lay people 60%. Both extracts were produced by female actors, but impressionistically the two extracts are rather different. Extract AA contained a lot of speech material relative to other extracts (14 syllables of speech material). The presence of more material, and therefore more cues on which listeners might have based their decisions, may have led to a greater increase in listeners' accuracy.

In contrast, extract AP is punctuated by two lengthy unfilled pauses (1.06s and 1.52s) due to the actor's confusion as to whether it was her or the operator's turn to speak. On realising it was her turn, she spoke the word 'argh' with extreme creak several times rather than producing the scream or vocalisation that 'argh' could represent. The unnatural timing of the production and the lack of a realistic 'argh' vocalisation probably led to the listeners' ability to classify this as an acted production. Even so, a minority of the listeners did not consider this extract to have been acted, possibly due to the absurdity of the production. Following the experiment, a few listeners gave verbal comments about that particular extract, commenting that it was so strange it must have been real (i.e. produced by a real victim). Interestingly, and as is described in more detail in §7.3, this extract was frequently perceived as having been produced by a male speaker due to the female actor's low pitch (ranging from 52 - 198Hz for this extract, which was mainly comprised of creaky voiced episodes and a production of the word 'no' with a rise-fall contour starting at 177Hz, peaking at 198Hz, and falling to 127Hz).

7.1.8 Consistently correct authentic extracts

Extract K, using speech material from the male victim in Case B, had the second-highest correct identification score of all the extracts, with 83% of responses being a correct victim attribution. 95% of all lay people correctly attributed the extract to a victim, versus 67% of forensic practitioners. The police call takers rated the speaker in extract K as a victim in 79% of their responses. Extract I, again using speech from the male victim from Case B, has another high correct identification score, with 78% of all responses correctly identified as 'victim'. The police call takers are most successful group in correctly classifying the extract (86%). Forensic practitioners have a correct identification rate of 67%, and lay people attain 80%. Both of extracts

K and I were produced by the same speaker, and both contained speech material rather than just screams. The higher accuracy of listeners' responses to this extract may be due in part to semantic content - much reference was made to the victim's physical injuries and attack during his call to the emergency services - and also due to his regional West Midlands accent. Although Actors 3 and 4 produced extracts with similar semantic content, both were speakers of SSBE. The extracts were presented in random order, but could have been recognised consciously or even unconsciously as having been produced by the same speaker. A tentative link may be drawn with perceptual accent studies, where regional accents are found to receive more positive evaluations in terms of personal integrity, such as being perceived as more sincere, than standard accents (Edwards & Jacobsen (1987), Coupland & Bishop (2007)). It could also just be that this speaker conveys his distress in a way that is perceptible and universally accepted, and recognised as authentic by other listeners for reasons that have yet to be ascertained.

7.1.9 Consistently incorrect acted extracts

Extract AI, containing speech produced by a female actor (Actor 10) from Case E, was incorrectly identified in all but 4% of responses. Only 7% of police call taker responses and 8% of forensic practitioner responses correctly identified the extract as having been produced by an actor. All lay people perceived it as having been produced by a victim. Extract AL, containing speech (though some of it can be categorised as 'other' using the distress taxonomy) from a female actor (Actor 12) using Case F material, was also frequently incorrectly identified as the speech of a victim (87% of all responses rated it as 'victim'). All of the police call takers believed the extract to have been produced by a victim, whereas 75% of the forensic practitioners rated it thus. Lay people misattributed it in 85% of their responses. Extract AC, containing speech from a female actor (Actor 8) performing Case D material, also had low scores, with only a 20% correct identification rate. 25% of lay people and 25% of forensic practitioners correctly identified the extract as having been produced by an actor, compared with 7% by the police call takers. All three extracts were produced by different female actors re-enacting different case material. The duration of the extract and the amount of clear, intelligible speech material vary across the extracts. However, common to the extracts is an impression of sobbing or

whimpering, which may have influenced listeners to rate them as having been produced by a real victim.

7.1.10 Consistently incorrect victim extracts

Victim extracts that were often incorrectly identified were Extract D, containing a scream from the male victim in Case A, and Extract AM, containing a scream from the female victim in Case F. Extract D was identified correctly in 20% of all responses, with identification rates of 20% for lay people, 14% for police call takers, and 25% for forensic practitioners. Extract AM was correctly identified in 28% of all responses, with a correct identification rate of 25% among lay people, 14% for police call taker responses, and 50% for forensic practitioners. These extracts, although produced by victims of opposite sexes, each contain a high-pitched screamed production. Devoid of context, some listeners made comments on the response sheet stating that they doubted the extracts as being produced by a human adult (some listeners thought the extract sounded like cat vocalisations or infant cries). This could explain why listeners were reluctant to mark it as having been produced by a real victim (though note that the forensic practitioners had a higher rate of correct identification than the other groups, suggesting that they took the strangeness of the extracts in their stride).

7.1.11 Individual extract analysis

To investigate further how individual extracts were perceived, average scores for each extract across participants, as well as variation within the responses, are provided in Table 7-3. It shows the overall scores and standard deviations of all responses together (excluding control extracts), as well as per participant group.

Table 7-3: Mean scores and standard deviations of all responses (excluding controls) to Q1 of the perceptual experiment. (LP = lay people, PC = police call takers, FP = forensic practitioners).

	Mean score of Q1 responses			S.D. of Q1 responses				
	LP	PC	FP	LP	PC	FP	Ave. mean	Ave. S.D.
Min.	1.75	1.71	2.25	0.41	0.55	0.45	2.04	0.63
Max.	4.20	4.21	4.08	1.34	1.50	1.23	4.17	1.27
Ave.	2.78	2.97	3.03	0.98	1.01	0.87	2.90	1.00

The forensic practitioners had the highest minimum score (2.25) and the lowest maximum (4.08) of all the groups, showing that their responses tended to be nearer the middle (the 'no decision' category), which is in line with their greater use of this category. Both the lay people and police call taker groups has similar minimum and maximum scores, though the lay people had a slightly lower average score (2.78) than the police call takers (2.97). The lay people therefore tend to respond with a score indicative of their belief that the extract was produced by a victim, whereas the police call takers appear to have a more balanced spread of victim- and actor-attributions scores.

The standard deviation of mean scores is lowest for the forensic group (0.87), which suggests that they are more consistent as a group in their categorisations of responses than are the lay people and police call takers. The average standard deviations for both these groups are similar (0.98 for the lay people and 1.01 for the police call takers), though the police call takers have the highest minimum and maximum standard deviations. Consequently, the police call takers are the most variable as a group when categorising the extracts.

For all groups, the standard deviation of the overall minimum mean score for extracts is lower than the maximum mean score. This could indicate more consistency across participants when they were rating extracts as having been produced by a victim.

Figure 7-7 and Figure 7-8 illustrate the mean and standard deviation of individual (non-control) extracts as grouped by the different participant types. Figure 7-7 shows that, on the whole, the lay people group tend to have a lower mean extract number (i.e. they rate extracts towards the victim end of the scale more than do the other groups), and Figure 7-8 shows that the forensic practitioners typically have a lower standard deviation score for each extract, thus showing that they rate more consistently as a group.

Furthermore, there are isolated points of interest illustrated in Figure 7-7 and Figure 7-8 which do not form part of a general trend. For example, in Figure 7-7 it can be seen that for extract O, which is a recording of Actor 6 screaming (based on material from Case C), each of the three participant groups rate very differently. The mean score for lay people for this extract is 3.1 (i.e. no decision) which may mean they were all unable to make up their mind and therefore scored near 3, or that they were willing to rate the extract, but some judged in favour of one direction, and others the opposite direction. The mean score for forensic practitioners was lower (scoring 2.3), indicating that - all things being equal - the extract was considered to have been produced by a victim, whereas the police call takers had a higher mean (scoring an extract average of 4.0) thus favouring (correctly) an acted production instead. The standard deviation scores associated with these means show that the police call takers had the lowest standard deviation for this extract (0.78), and therefore on a group level seem to be consistently rating this extract as acted, whereas the other two groups have higher standard deviations, suggesting that the mean of means is composed of a mixture of scores corresponding to both acted and authentic attribution ratings. This extract was impressionistically very similar to the authentic version of the extract.

Figure 7-7: Mean scores for each extract (ordered in overall ascending mean). The position of X indicates whether the extract was produced by a victim (1) or an actor (5).

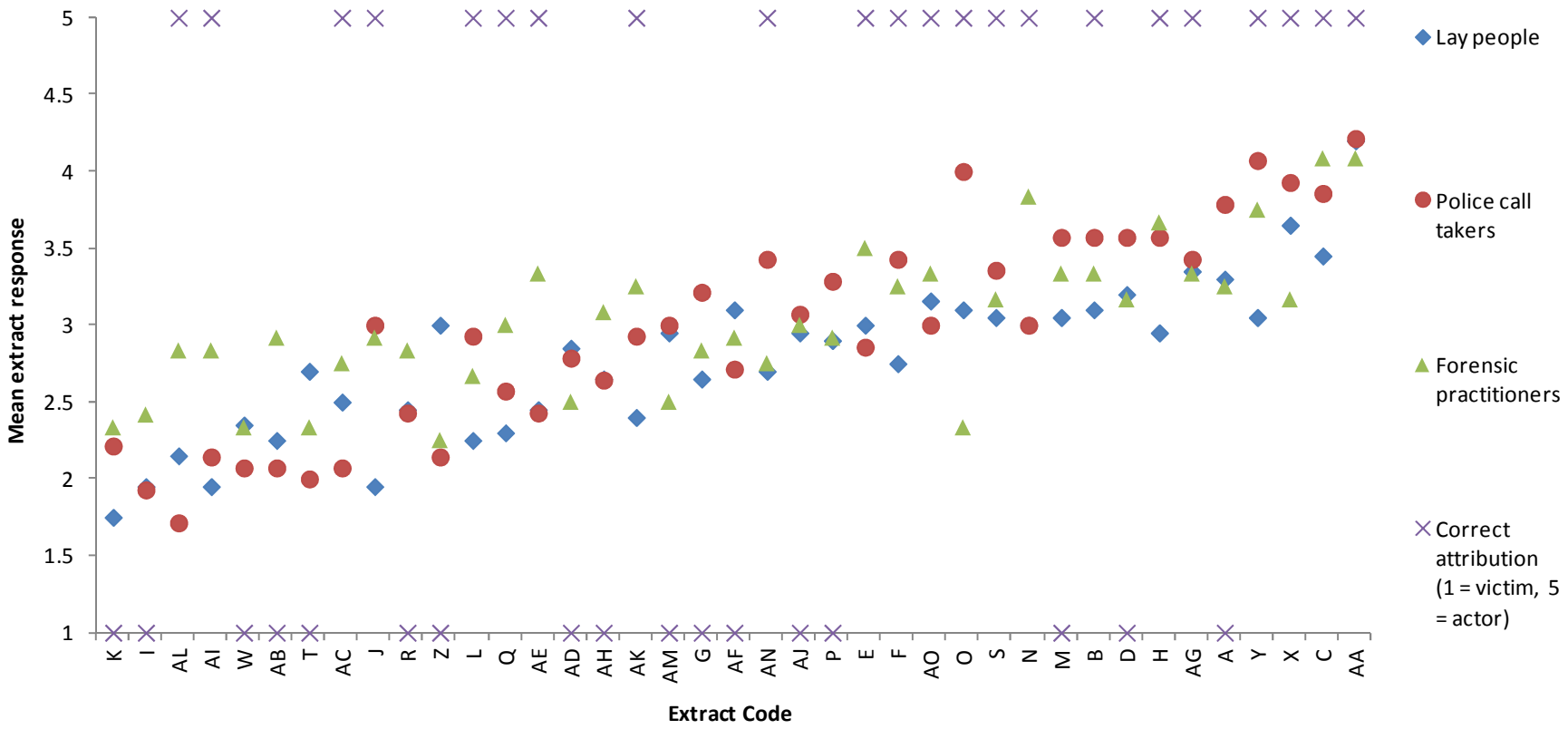
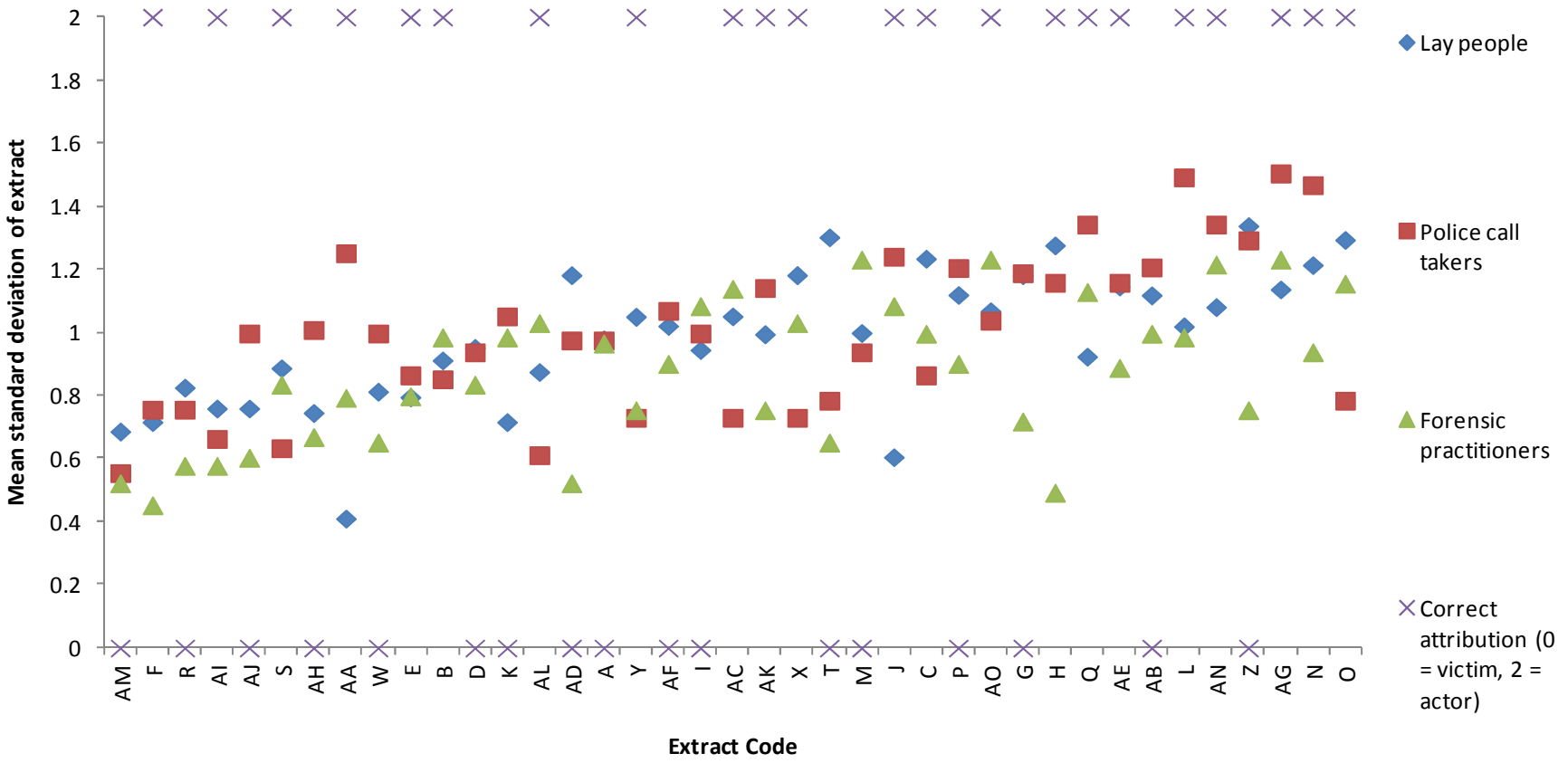


Figure 7-8: Mean standard deviation for each extract (in ascending order of overall standard deviation). X indicates whether the extract was produced by a victim (1) or an actor (5).



On the other hand, Extract AJ, containing speech from the female victim in Case F, and Extract AG, an extract of a female actor (Actor 9) producing speech based on Case D, have very similar means, at 3.0 and 3.37 respectively. However, their standard deviations reveal that the police call takers are the least consistent group, as they have the highest standard deviation (1.00 and 1.50 for each extract, respectively) indicating that members of the group may be rating at opposite ends of the scale, whereas the other two groups are more similar and less variable in their responses (0.76 and 0.60 for AJ, and 1.14 and 1.23 for AG), showing that both are more likely to rate near the centre of the categorisation scale. Of the two extracts, AJ shows more consistency with respect to standard deviation scores than AG, indicating that listeners were more in agreement when rating AJ.

Extract AM has the lowest overall standard deviation for an extract (0.63 across all participants), with very similar standard deviations for each of the three groups (0.69 for lay people, 0.55 for police call takers, and 0.52 for forensic practitioners). Each group was therefore consistent in its categorisation of this extract, though the categorisation of the extract is different for each group. Both police call takers and lay people rated near the centre of the categorisation scale, i.e. 'no decision' or weak ratings in both directions, whereas the forensic practitioners were inclined to rate (correctly) the extract as having been produced by a victim.

In addition, Figure 7-7 and Figure 7-8 highlight in which direction the correct identification of the extract should lean. With the mean scores presented in ascending order (Figure 7-7), it can be seen that those on the right of the diagram trend towards an actor identification, and that for most of the extracts this is indeed true. Therefore, actor extracts are typically correctly identified. By contrast, those extracts on the left of the diagram, with lower mean scores suggesting a victim identification, are not always correctly identified. There appears to be a mix of both acted and real extracts that are judged to have been produced by a victim. This supports the previous conclusion that misattributions are typically rated as 'victim' when they are in fact produced by an actor, as illustrated earlier in Figure 7-3.

The overall mean extract score did not significantly change across the three participant groups (χ^2 (2) = 0.94, $p > 0.5$). Furthermore, overall mean extract scores for each participant group (as well as all three populations as a whole) did not differ significantly with regard to the type of extract (screamed productions vs. speech productions); the speaker of the extract (actor vs. victim); and the sex of the speaker (male vs. female).

7.2 Listeners' confidence level when distinguishing between acted and authentic distress

Question 2 of the experiment concerned how confident participants felt when assessing whether the extract was produced by an actor or a victim. Participants made use of a 5-point non-numerical scale and were asked to choose between *very certain, quite certain, neither certain nor uncertain, quite uncertain* and *very uncertain* for each extract. Each option was coded as a numerical value so that it could be converted into a percentage, and to enable statistical analyses (Table 7-4).

Table 7-4: Coded responses used to enable statistical analysis of Question 2 of the experiment.

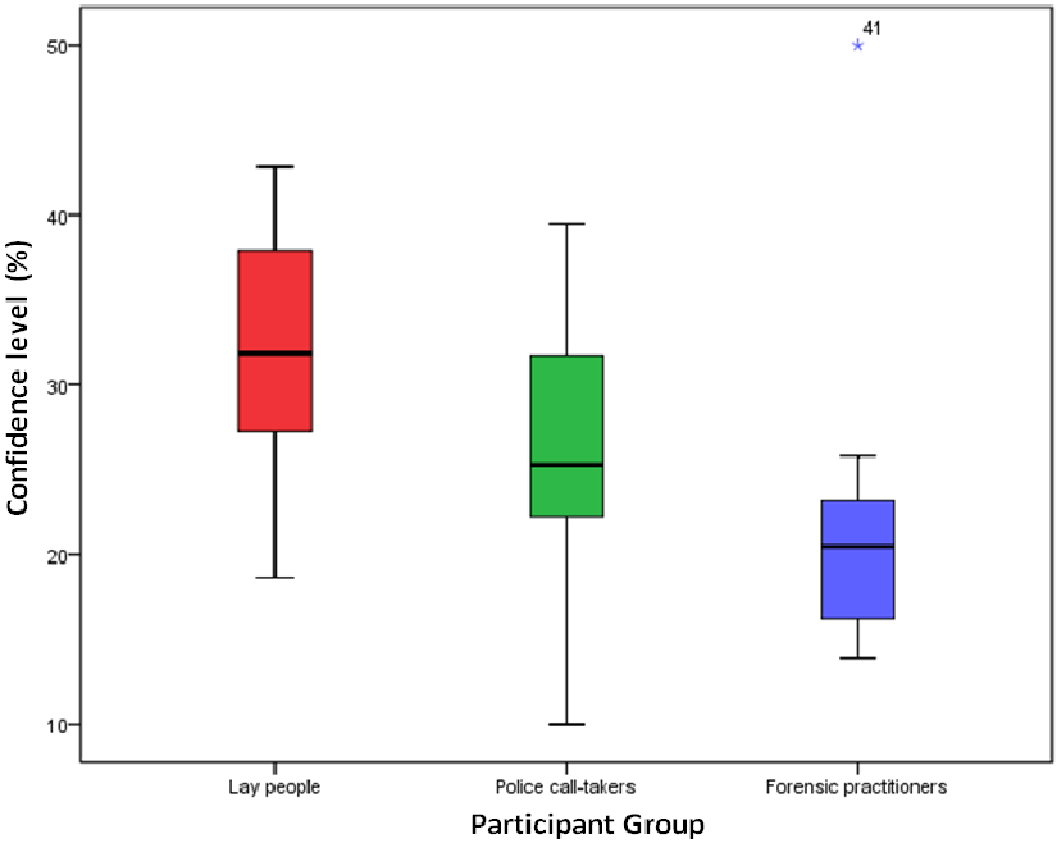
Options on response sheet	Very certain	Quite certain	Neither certain nor uncertain	Quite uncertain	Very Uncertain
Coding for statistical analysis	100	75	50	25	0

Figure 7-9 illustrates the variation in confidence levels across the listener groups. Lay people expressed the most confidence in their responses (median = 31.84) and forensic practitioners the least (median = 20.47), though note the general low levels of confidence expressed by both groups. Police call takers had the greatest range of confidence levels out of all the groups. This group included the least confident individual (Median = 25.29, range = 29.47, min. = 10.00). The forensic practitioners had the smallest range of confidence levels (range = 14.24), though there was also an outlier in this group - one forensic practitioner rated much more confidently than any other participant (averaging 50% confidence level across all extracts). This participant also had the lowest correct identification rating in his/her group for

Question 1 of the experiment (mean correct identification rate including 'no decision' responses was 26%).

A Kruskal-Wallis test was performed. It showed that confidence level was significantly affected by participant group membership ($H(2) = 13.32$, $p < 0.001$). Jonckheere's test revealed a significant trend in the data and one that is well illustrated in Figure 7-9: confidence levels decreased from lay people to police call takers to forensic practitioners ($J = 159$, $z = -3.77$, $r = -0.56$).

Figure 7-9: Min., max., median, and interquantile ranges of confidence levels across participants.



Kolmogorov-Smirnov Z tests revealed that there were no statistically significant differences between the confidence levels of those who completed experiment A and those that completed experiment B across the three participant groups, nor between confidence levels among older and younger participants across the three groups.

Overall confidence levels among male participants in police call taker and forensic practitioner participant groups (median = 27.60 and 17.50, respectively) did not significantly differ from those of female participants (median = 22.50 and 22.58, respectively), Kolmogorov-Smirnov $Z = 0.97$, ns, $r = 0.26$ (police call takers) and Kolmogorov-Smirnov $Z = 0.83$, ns, $r = 0.24$ (forensic practitioners). However, female lay participants reported significantly higher confidence levels (median = 37.50) than their male counterparts (median = 30.30), Kolmogorov-Smirnov $Z = 1.42$, $p < 0.02$, $r = 0.32$.

A four-way mixed design ANOVA was also performed in order to test for significant differences caused by characteristics of the extract (as part of a by-items analysis), across the three participant groups. It confirms the result from the Kruskal-Wallis test on the previous page. However, the ANOVA is calculated using extract data, whereas the Kruskal-Wallis is based on participant data. All effects are reported as significant where $p < 0.05$. There was a significant main effect of participant group on the overall confidence level of the experiment extracts (minus the eight control extracts) (F (2, 62) = 115.40). Contrasts confirmed that the lay people had a significantly lower average confidence level across all extracts than did their police counterparts (F (1, 31) = 35.98, $r = 0.73$), but had a significantly higher average than the forensic practitioners (F(1, 31) = 91.57, $r = 0.86$).

There was a significant interaction between participant group and whether the extract contained speech or screamed productions (F(2, 62) = 3.67). ANOVA planned contrasts revealed no significant interaction between lay people and police call takers (F(1,31) < 1, $r = 0.09$), but there was a significant interaction between lay people and forensic practitioners (F(1, 31) = 4.37, $r = 0.35$). The interaction graph in

Figure 7-10 shows that confidence levels were lowest among forensic practitioners for both screamed and speech productions, but they fell even lower for speech productions. In contrast, confidence levels among lay people, although already quite high, become higher still in extracts containing speech productions.

In addition, there was a significant interaction between participant group and whether the voice in the extract was produced by a male or a female speaker ($F(2,62) = 3.51$). ANOVA planned contrasts again revealed no significant interaction between lay people and police call takers (F (1, 31) <1, r = 0.05), but there was an interaction between lay people and forensic practitioners (F (1, 31) = 4.76, r = 0.36). In a similar vein to the previous interaction, the interaction graph shows that confidence levels are lowest amongst forensic practitioners for all extracts, but are even lower among this group for extracts produced by female speakers, whereas confidence levels among lay people, which are again already higher, increase further for extracts produced by female speakers. Lay people express more confidence when rating female voices, whereas forensic practitioners are more confident when rating male voices.

Figure 7-10: Confidence ratings across all extracts according to participant group and type of production in extract.

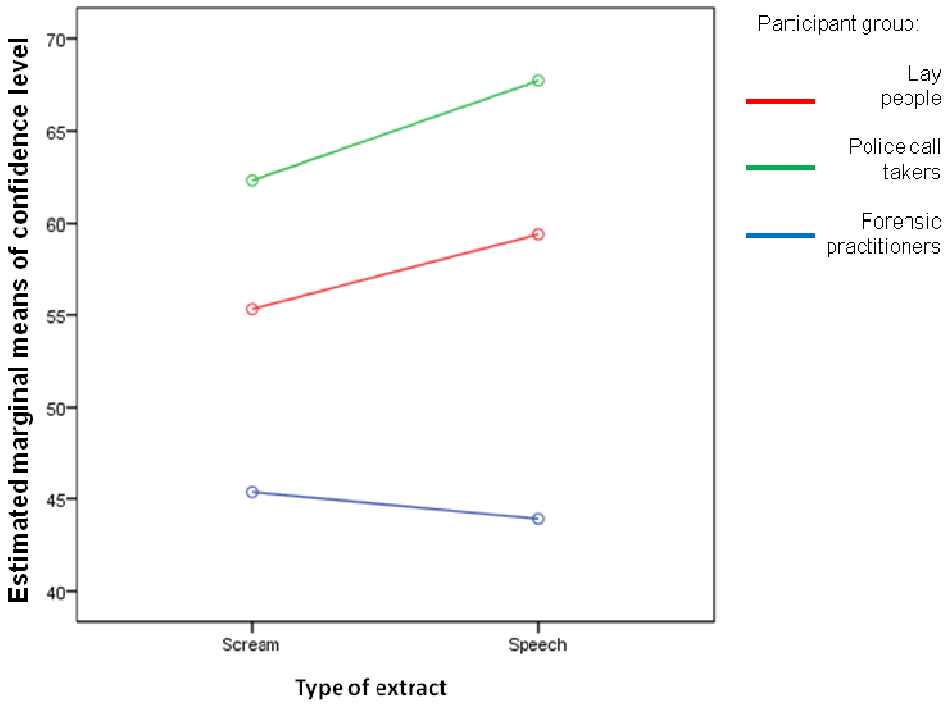
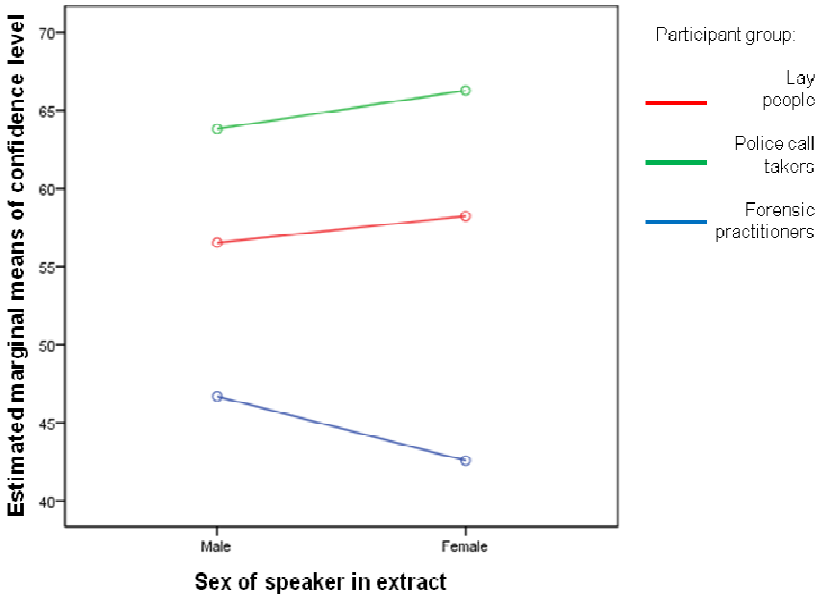


Figure 7-11: Confidence ratings across all extracts according to participant group and sex of listener.



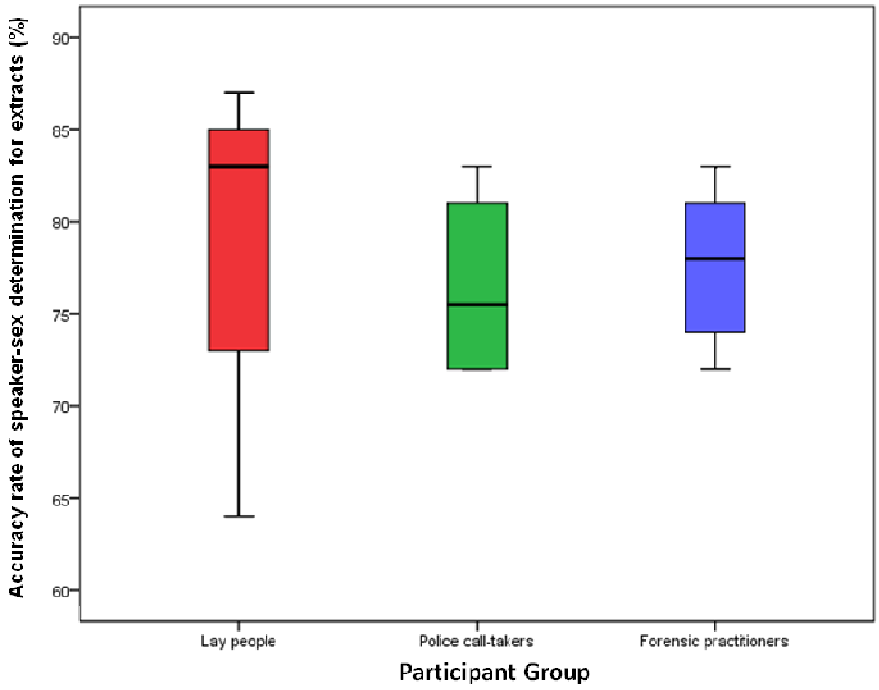
One additional question is whether confidence levels corresponded to accuracy levels. In this context it is worth noting that while the survey question allowed participants to distinguish between 'probably victim' and 'definitely victim', and 'probably actor' and 'definitely actor', this distinction was not taken into account in gauging the accuracy level of each participant for this analysis, nor are 'no decision' responses included. Therefore, the accuracy level here is independent of confidence level. A bivariate correlation showed that the participants' confidence level was significantly inversely related to correct identification rate ($\tau = -0.60$, $p < 0.001$) in that as confidence levels decreased, accuracy increased. However, if analysed by group, this trend holds for lay people ($\tau = -0.56$, $p < 0.001$) and police call takers ($\tau = -0.50$, $p < 0.05$), but no such trend is statistically significant for forensic practitioners ($\tau = -0.31$, $p > 0.05$).

7.3 Listeners' differentiation of male and female voices in distress

The third question in the experiment asked listeners to identify whether the voice they heard in the extract was that of a male or a female. Figure 7-12 shows that the median score for accurate speaker-sex determination was at least 75% across all participant groups. The lay people group contained individuals with both the best and worst correct rate of speaker sex-identification, and this group had the highest median overall (83%). Police call takers and forensic practitioners had smaller

interquartile ranges, meaning that the individuals of which these groups were comprised performed neither worse nor better than the lay people, but that they attributed sex to speakers more consistently than did the lay people. The forensic practitioners had a slightly higher median (78%) than the police call takers (76%). A Kruskal-Wallis test revealed no statistically significant differences between the three groups. A series of Mann-Whitney tests showed that there were no statistically significant changes between participants doing experiment A vs. experiment B, or male participants vs. female participants, and that level of education was not related to the scores for lay people or police call takers.[18] However, correct speaker-sex scores among old participants (median = 74.00) were found to be significantly lower than those for young participants ((median = 82.00), Kolmogorov-Smirnov $Z = 2.03$, $p < 0.001$, $r = 0.30$).

Figure 7-12: Min., max., median, and interquartile ranges of correct speaker-sex determination extracts across participant groups.

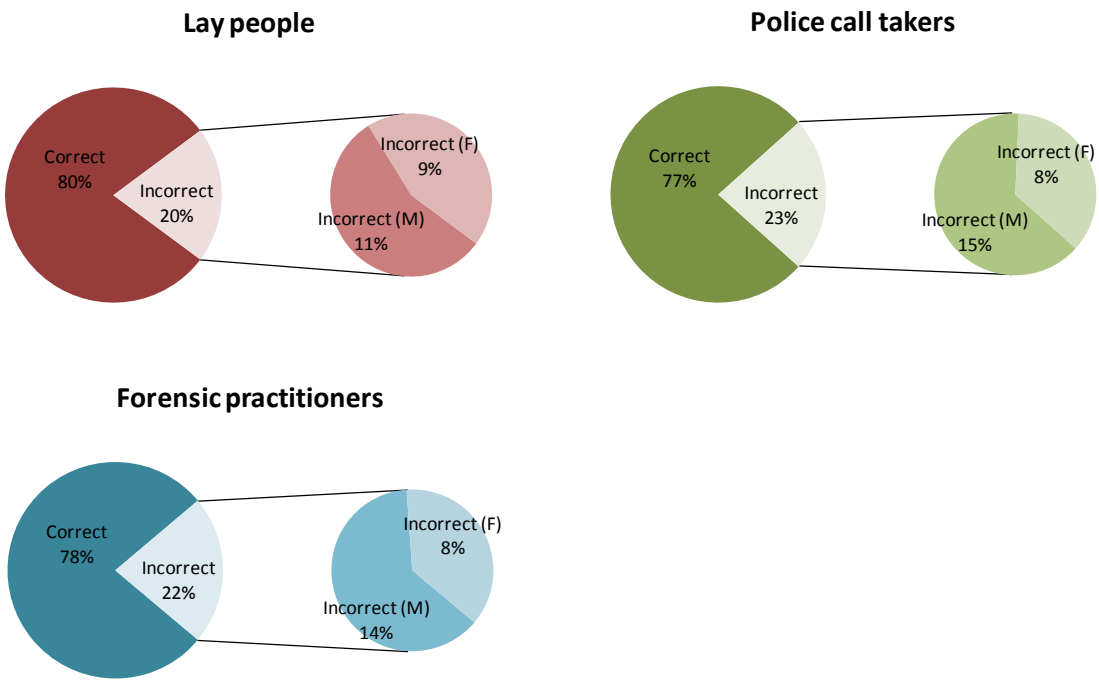


The direction of incorrect speaker-sex responses is illustrated in the pie charts in Figure 7-13. Mean correct and incorrect scores are similar across all groups, though in all cases the majority of incorrect responses occurred when the speaker of the

[18] Forensic practitioners were omitted from this analysis, as all were educated to postgraduate level.

extract was male. For police call takers, incorrect determination of speaker sex in the extracts was significantly higher when the speaker was male (median = 14.89) than female ((median = 7.45) , z = -2.99, p < 0.01, r = -0.80). A similar pattern was observed among the forensic practitioners; incorrect determination of speaker sex in stimuli was significant higher for male speakers (median = 14.89) than female speakers ((median = 8.51), z = -3.07, p < 0.01 (0.002), r = -0.89). Although lay people also displayed a higher rate of incorrect sex identifications for male speakers (median = 10.64) than for female speakers (median = 7.45), this result was not significant.

Figure 7-13: Pie charts showing the nature of incorrect speaker-sex identifications across all participant groups.



There was no correlation between the rate of correct sex determination and correct identification rate (τ = 0.11, p > 0.05). Therefore, there appears to be no relationship between accuracy in one variable and accuracy in the other.

A by-items analysis of the data revealed that the majority of extracts generated high levels of accuracy in speaker-sex identification across the groups; 33 out of 43 extracts had a correct sex determination rating in $\geq$ 80% of all responses (Figure 7-14). Seven had poor accuracy rates, with less than 40% of responses being correct

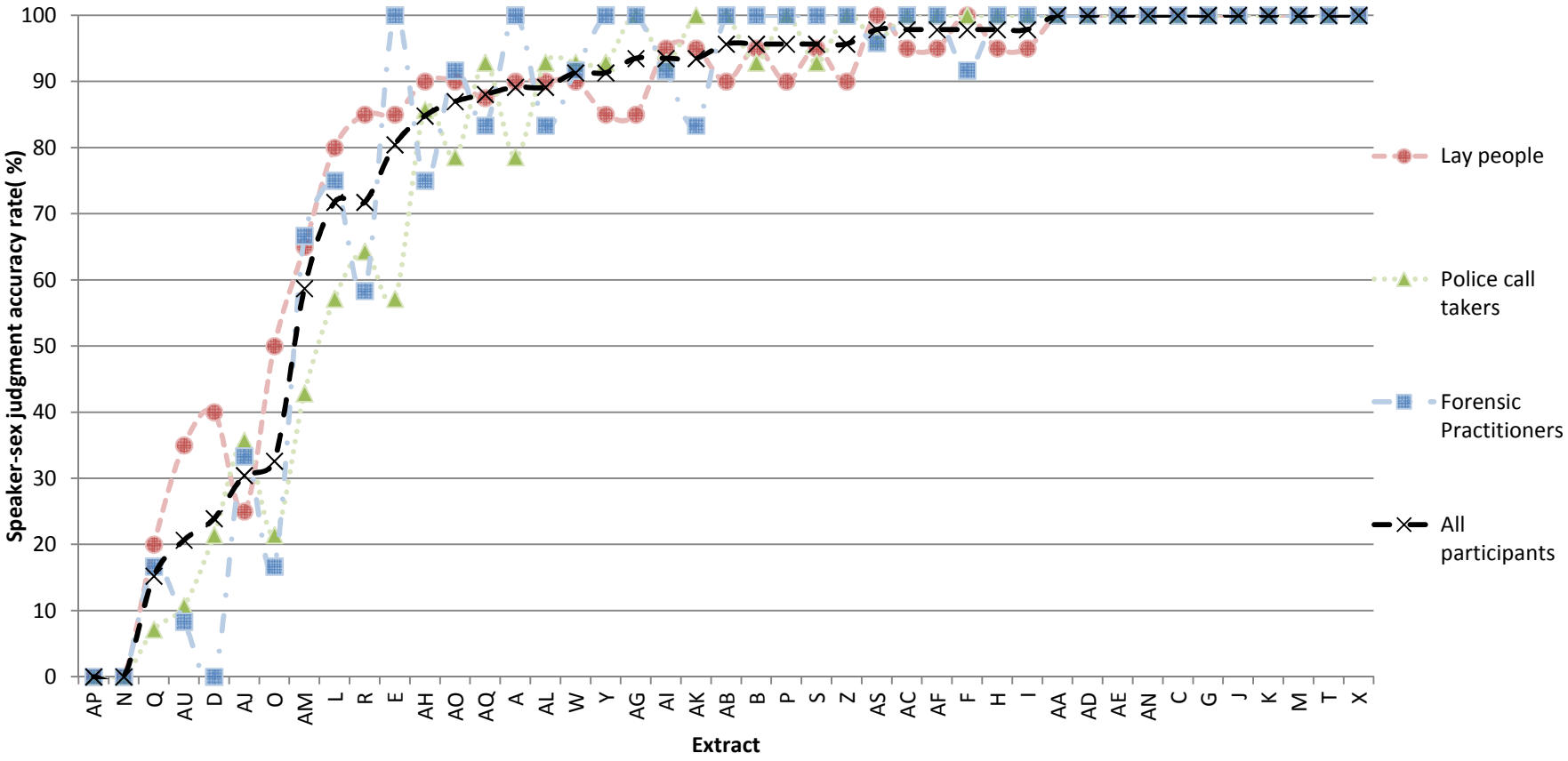
for sex identification. Of these, extracts AP and NP were never correctly identified. Extract AP, which is described in more detail in §0, contained productions by Actor 12 using material from Case F. This extract has 3 distinct sections. The first section contains an open vowel-like production which at first contains intermittent creak and periodicity (ranging from 52 - 75Hz), becoming more modal in a "moan" section (averaging 118Hz). Following a pause of 1.06 seconds, 'no' is produced with a mean F0 of 168Hz, with a rise-fall F0 contour starting at 177Hz, peaking at 198Hz, and falling to 127Hz. Another pause follows, lasting 1.52 seconds, and then 895 milliseconds of creak that sounds almost like belching can be heard. The fact that this extract is frequently misidentified is not surprising given that the cues we would typically expect of a female voice, including higher F0 values than these, are not present. A few listeners commented that the unnaturalness of the timing of the extract, i.e. the unfilled pauses lasting over a second, might also confuse listeners and direct their attention away from the sex of the speaker. Extract N, a male actor (Actor 5) vocalising and screaming material from Case C, with some "sobbing" quality reminiscent of extracts AC, AI, and AL (see §7.1.9) has a high F0 (mean = 364 Hz, min. = 279 Hz, and max. = 440 Hz), presumably priming listeners to perceive the vocalisations as female. Similarly, extracts D and O contain screams produced by male individuals. Extract D contains a scream produced by Victim A. The F0 of the onset of the scream is 909Hz and the end of scream falls to 527Hz (mean = 831 Hz, min. = 527Hz, max. = 926 Hz). Extract O contains a series of four screams from Actor 6 (mean = 757 Hz, min. = 407Hz, max. = 892 Hz). Extracts Q and AU do not contain screams, but speech from male actors with high F0. Extract Q contains impressionistically "hurried" speech from Actor 5 (mean = 242 Hz, min. = 181 Hz, max. = 305 Hz). Extract AU contains speech with elements of sobbing (mean = 337 Hz, min. = 196 Hz, max. = 443 Hz).

Interestingly, not all screamed extracts produced by males were misinterpreted. Extract M from Victim C was always identified correctly as having been produced by a male speaker, despite it containing screamed productions averaging 325 Hz (min. = 148 Hz, max. = 482 Hz). One noticeable difference separates this extract from those that were frequently not perceived as having been produced by males: the

minimum Hz value, i.e. the lowest part of the range, is still typical of male speaker, even though the speaker's range reaches that typical of a female speaker.

The only other extract produced by a female speaker that received a poor speaker-sex accuracy rate was extract AJ. It contains a vocalisation produced by Victim F (mean = 431 Hz, min. = 297 Hz, max. = 510 Hz). The vocalisation has a high F0 throughout. It is not clear which features are absent from the extract than might cue femaleness, nor is it clear which features might be present to cue maleness.

Figure 7-14: Line chart showing average accuracy judgments of speaker sex across all listeners ranked by grand mean.



7.4 Listeners' written comments

The fourth section of the experiment response sheet invited listeners to make any notes or comments about the extracts they heard. A table displaying all the comments is provided in Appendix G5. Overall, there was a considerable amount of variation in the number of written comments contributed by the various participants, with most participants failing to comment altogether, and with the maximum number of comments contributed by a single participant being 35. Table 7-5 shows the number of comments that were contributed by different participants, according to the participant group:

Table 7-5: The number of written comments volunteered across participant groups.

	No Comments	1-5 comments	6-15 comments	16-30 comments	31+ comments
Forensic practitioners	2	3	3	2	2
Lay people	13	5	1	1	1
Police call takers	5	7	2	0	0

As illustrated in the table, forensic practitioners were by far the most verbose, while police call takers were more selective in their commenting. Their reluctance in making notes may be due to the fact that there were participating in the experiment during the work shift. Although they had been given permission to take part in the experiment, it appeared that the majority were willing to answer the first three questions for each extract, but were eager to avoid spending additional time writing comments and keen to return to work.

The comments can be roughly categorised into three types. The first type of comment, to which roughly 70% of the written comments belonged, was explanations of the scores given by participants. As is perhaps not surprising, the type of explanation varied with participants' experience in phonetics. Forensic practitioners tended to provide technical assessments ("high intelligibility/articulateness a significant cue", "judgment largely based upon voice quality") and lay people and police call takers tended to be more vague ("Weirdly calm", "the scream sounded misplaced somehow", "Sounded in need of help"). Overall, there was very little consistency to the reasons given by participants for

their judgments of any extract, and it is difficult to draw any useful conclusions from these comments.

The second type of comment concerned the difficulty of the task, in particular when extracts were considered too short to allow a decision to be made. Extracts AH, AI, AJ, S and W were all judged to be too brief by the majority of participants who commented on them. What is of potential interest here is that this concern was exclusive to forensic practitioners and police call takers, with lay people never directly commenting on the length of an extract. It is impossible to determine based on the current data, whether this means that lay people are more willing to make judgments about short audio extracts in general, or whether they are simply less willing to challenge the design of an experiment. This is a question for potential future research, as is the question of how long an extract may need to be for more experienced people to feel comfortable judging it.

The final type of comment consisted of participants providing an additional assessment of the extract that was not directly relevant to their task (for example 'Northern', or 'Elderly'). These were infrequent, were contributed in roughly equal amounts by all three participant groups, and were distributed randomly across the various extracts, and it is difficult to assign any meaning to them.

There were some individual patterns that could be observed for specific participants. For example, participant 2 (a forensic practitioner) seemed to start out being very doubtful of his/her own ability to make useful judgments, but grew more confident in later extracts, while participant 214 (a police call taker) made several guesses as to the age of the victim/actor (perhaps because this is information that is often relevant in his/her line of work). However, there were no cases where a participant's comments gave any indication that his/her data might have been problematic in any way.

7.5 Chapter summary

The foregoing chapter shows that when attempting to distinguish between authentic and acted distress using short audio extracts, lay people give the smallest number of

correct responses and the highest number of incorrect responses, i.e. they are the listener group that performs most poorly in the perceptual experiment. The police call takers and the forensic practitioners, i.e. the listener groups with familiarity with authentic data, are both equally good. Forensic practitioners give the smallest number of incorrect responses, but the highest number of 'no decision' responses. Their higher rate of 'no decision' responses did not improve their correct response rate, but it did reduce their incorrect response rate. Listeners tended to express low levels of confidence when assessing both authentic and acted distress, but lay people, especially females, are the most confident group on the whole. Forensic practitioners are the least confident group, presumably as a result of their training rather than their competence. Listeners performed well in correctly identifying the sex of the speaker in each extract, though some extracts produced by males with high F0 values were sometimes misidentified as having been produced by females. There were no significant differences between listener groups when assessing the sex of the speaker in the extract. A cluster analysis revealed that the participant groupings of lay people, police call takers, and forensic practitioners were partially well-motivated since the clusters generated independently in the analysis corresponded to experience level, and separated the lay listeners from the police call takers and forensic practitioners. Finally, where listeners provided written comments about the extracts they heard, the comments varied in number and content. Forensic practitioners tended to provide technical assessments, whereas lay people often wrote rather vague comments. Both the forensic practitioner and police call taker listener groups made frequent reference to some extracts as being too brief to form a judgment, whereas lay people expressed no such concern.

8. Discussion

This chapter considers the implications of the results reported in chapters 5 and 7 for the field of forensic speech science and phonetics generally, in terms of the research questions presented at the beginning of this thesis. To recap, the investigation centred on two key questions. Firstly, to what extent can specific acoustic measures be used to classify distress speech? Secondly, can listeners perceive the difference between authentic and acted distress?

To attempt to answer the first question, the thesis investigated vocal cues that may be used to identify distress, and examined the material to see whether there were any cues that distinguished acted from authentic distress responses. In this chapter, the first part of the discussion focuses on these questions, specifically considering the special cases of F0 and intensity in distress productions, and exploring vocal correlates of distress generally as well as means to distinguish between acted and authentic distress.

For the second research question, the thesis explored listeners' accuracy and confidence when comparing acted with authentic distress, and considered whether familiarity with forensic data resulted in increased accuracy and/or confidence. A further sub-question issuing from the second research question concerned whether listeners can differentiate between male and female distress responses. Listeners' perceptions were addressed in the second part of the discussion, which explored perceptual differences between acted and authentic distress at first, and then focused on the accuracy of the listeners.

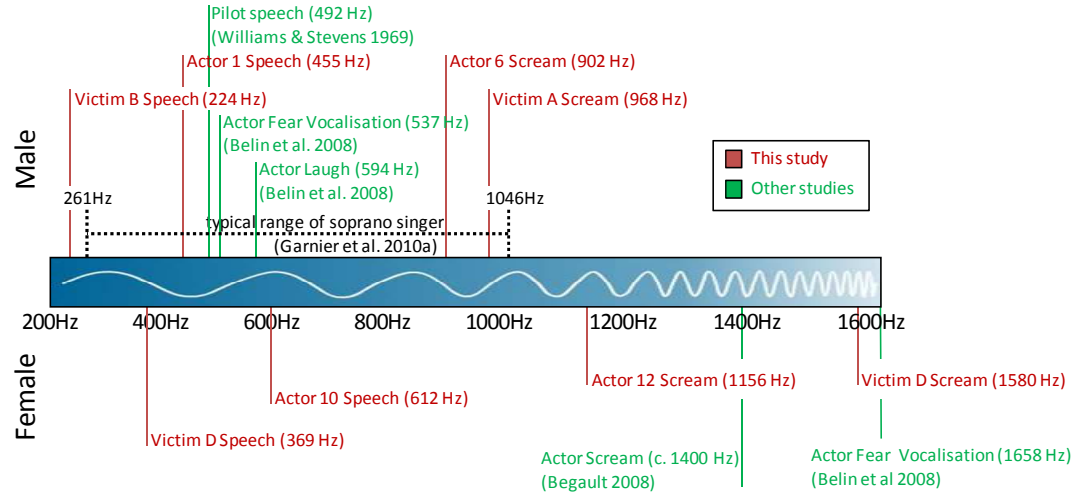
8.1 Acoustic cues to distress

8.1.1 The special cases of F0 and intensity in distress productions

Chapter 5 revealed that individuals in distress often displayed differences in the acoustic signal when their distress recordings were compared to their reference (non-distress) material. As found in previous affective speech studies, the most salient changes concerned the parameter of F0. It was found that the maximum F0 observed in this study exceeded F0 values typically reported for emotionally aroused speech, reaching just under 1000 Hz among male victims (on a par with the top range of a

typical soprano) and 1600 Hz among female victims. As reported in the literature review (§2.6.1), an increase in F0 during emotionally aroused speech is expected, but perhaps not to this extent. Only two previous speech studies report results close to these. Firstly, the Montreal Affective Voices (MAV) dataset reports comparable values for acted emotional speech in females, up to 1658 Hz, though not in males (Belin et al. 2008). This dataset contains productions that are most similar to the data in this study in that it incorporates non-verbal emotional material. Secondly, Williams & Stevens (1969) document the speech of a pilot in a dangerous flying situation reaching 492 Hz. A visual representation of pitch maxima across speech studies (including professionally trained voices) is provided below.

Figure 8-1: A comparison of pitch maxima across speech studies (including this investigation).



The onset of increased F0 is rapid in both victims and actors and so adjustments to the individual's vocal apparatus, e.g. tension, airflow, larynx configuration and musculature, and the vocal folds themselves, are made in just milliseconds. The larynx is therefore capable of extreme rates of vocal fold vibration and extreme tension of the folds for at least short durations when responding to either a genuine threat or a simulated threat. It is of interest to discuss how and why individuals, both trained and untrained in vocal techniques, demonstrate such extreme increases in F0.

Firstly, trained singers typically receive vocal training and performwarm ups in order to minimise any potential damage and to optimise their singing power and range, e.g. Husler & Rodd-Marlin (1965: 35), David (1995: 98). It is reported that singers found

"it was easier to sing, particularly at high pitches" following vocal warm up exercises (Elliot et al. 1995: 39), though note that this is not always corroborated acoustically (Motel et al. 2003). The typical range of a soprano opera soloist is from C4 (261Hz) to C6 (1046 Hz) (Garnier et al. 2010a), though coloratura sopranos, singers specialising in high, elaborate melodies, may sing at considerably higher pitches, e.g. from D6 (1175 Hz) to beyond C8 (4186 Hz) (Garnier et al. 2010b). In opera, one of the most demanding performances for a soprano is found in the Queen of the Night aria from Mozart's *The Magic Flute*, as it contains some notes on F6 (1397 Hz) (Garnier et al. 2010a). The use of extremely high pitches is not limited to classical singing, but can also be found in the repertoire of female jazz and pop singers such as Mariah Carey and Georgia Brown who often sing above C6 (Garnier et al. 2010a). In addition, grunts, screams, rattles, and growls, collectively referred to as Extreme Vocal Effects (EVEs), have become more popular in non-classical music throughout the past century, particularly from the 1970s onwards with the popularity of rock and heavy metal music, and the advent of hardcore music (e.g. Black Flag), and grindcore music (e.g. Napalm Death) (Nieto Caballero 2008). Some female singers occasionally have perceptible epiglottal constriction as part of their performance, e.g. Alicia Keyes in the song *Falling*. Although little research has been conducted detailing the physiology of EVEs, it has been reported that a growing number of young people have been experiencing vocal disorders and polyps due to the popularity of EVEs in modern music (Nieto Caballero 2008: citing Van Onze (2007)). On the other hand, recent research suggests that EVEs can be produced without damage to vocal health if performed correctly(McGlashan et al. 2007).

In this study actors do not always produce higher F0 values than real-life victims (Figure 8-1), despite the former having received vocal training as part of their general dramatic training, and having also performed vocal warm up exercises in order to prevent vocal damage prior to being recorded. Roberts (Roberts 2010) found that mean F0 values produced by actors from popular film and TV were higher than those produced by real-life victims. A possible explanation for this is that the actors deliberately exaggerated F0 for the benefit of the audience because of it being a salient feature of aroused emotion. It should be borne in mind, however, that it was unknown whether the actors' speech productions had been modified post-production

for the benefit of the sound track of the film/TV programme. Nonetheless, it appears that in a genuine life-threatening situation, an individual without vocal training may undergo a hitherto ill-defined physical and/or mental reaction when in distress and quickly produce extremely high rates of focal fold vibration. This extreme F0 increase is not typically produced or required for day-to-day interactions, and is not necessarily attainable by stage and music performers, despite years of vocal training.

Secondly, the increase in F0 could be the result of a reflexive and (mediated) physiological response to danger, or an evolutionary response to a perceived threat which may include communicative meaning. On the one hand, changes to physiology when facing a stressful situation are produced as a survival mechanism, ultimately leading to a faster rate of respiration and an increase in muscle tension. This in turn is likely to result in a tensing of the vocal folds and consequently an increase in F0. However, although this could explain F0 increases in distress speech, it does not account for the extent of the increase that has been observed. Instead, variation in the F0 observations across individuals might relate to variation in the individual's evaluation of the threat, i.e. psychological factors mediate the physiological response, resulting in an individualisation of the response (Kirchhübel et al. 2011: 78). We can therefore ask whether individuals screaming with a high F0 beyond the upper threshold of their habitual pitch range evaluate the distress differently from those who do not.

On the other hand, an alternative explanation is that F0 increases in distress speech (partially) adhere to the 'frequency code' (Ohala 1984). Ohala argues that high F0 is associated with the traits of being small, non-threatening, submissive, and co-operative, whereas a low F0 is characteristic of threat, dominance, self-sufficiency, and intent to win in a contest. This communicative meaning is said to arise because the physiology of small members of species typically leads to a higher F0 (and conversely lower F0 for larger members of species). Animals can therefore either show off how large they are (and therefore be considered not worth fighting since the opponent is likely to lose) or how small they are (and therefore be viewed as non-threatening to the larger animal (Morton 1977)). In the life-threatening situations that invoke distress presented here, the victims with the greatest pitch maxima (Victims

A, D, F) are those who responded to an attacker who was still present, lending support to the hypothesis that their high pitched productions may have been an attempt to minimise further injury by appealing to the attacker as a weaker and non-threatening individual.

The sound symbolism of F0 in speech is thought to be consistent across cultures and across species, not only in terms of communication of emotional states, but also in facial expression and, in human cultures, vocabulary (Ohala 1984; Ohala 1996; Ohala 2009). Positioning the mouth with retracted lip corners (resembling smiling) is argued to represent submission in apes because this manoeuvre shortens the vocal tract, thus creating higher frequencies, whereas protruded lips (the 'o' face) signify aggression and disapproval, and doing so lengthens the vocal tract to produce lower frequencies (Ohala 1984: 6). The sound symbolism of mouth position and lip shape is exemplified in vocabulary across cultures by a prevalence of close front vowels in words signifying 'small', and close back vowels to signify 'large', across different languages (Table 8-1 and Table 8-2). The words denoting 'small' tend to contain segments with higher acoustic frequency (voiceless obstruents and vowels with a high F2 for vowels) than those denoting 'large' (voiced obstruents and vowels with a low F2). All of the example words denoting 'small' feature the front, high vowel [i], which admittedly usually features the lowest F1 in a speaker's repertoire. However, in the absence of F0, for example in whispered speech, it has been found that F2 corresponds to pitch, e.g. Thomas (1969), McGlone & Manning (1979). Although we might expect the 'small' and 'large' example words in Table 8-1 and Table 8-2 to be produced with F0, a question arises concerning the role of F2 in pitch perception. The high F2 in the 'small' words may be perceived as having a high F0, which would support Ohala's notion that high F0 can be associated with 'small' or 'weak'. The sound symbolism between stature and F0 may also explain why the Yoruba word [bírí], with high tones denotes 'small', yet [bìrì] with low tones denotes 'large'.

Table 8-1: Examples of words meaning 'small' (adapted from Ohala 1984, 1994).

Word/Morpheme	Language	Translation
[kítsíkítsí]	Ewe	'small'
[bírí]	Yoruba	'be small'
[tʃiko]	Spanish	'small'
[mikros]	Greek	'small'
[pətit]	French	'small'

Table 8-2: Examples of words meaning 'large' (adapted from Ohala 1984, 1994).

Word/Morpheme	Language	Translation
[g͡bàg͡bàg͡bà]	Ewe	'large'
[bìrì]	Yoruba	'be large'
[gordo]	Spanish	'fat'
[makros]	Greek	'large'
[gʁã]	French	'large'

Although facial expression and lip movement were not investigated formally in this research, the acted data were video recorded. It can be seen that productions involving high F0, namely screams, were typically produced with an open mouth and lowered jaw with no lip protrusion. Moreover, although there was some acoustic evidence to show that vowel articulations in distress speech were realised with an increase in F2 in actors' unrehearsed distress data, the principal finding was that if vowel articulations did change, it was generally in the direction of an increased F1 (possibly linked to the lowering of jaw in these productions). In these two respects, acted and authentic distress productions do not fully adhere to the sound symbolism principles advocated by the frequency code hypothesis.

The presence of non-linear phenomena in acted and authentic distress productions with high F0 further supports the idea of an evolutionary process, since non-linear phenomena are reported across other species. They are found in non-primate mammals, e.g. in wild dogs (Volodin & Volodin 2003) and kittens (Riede & Stolle-

Malorny 1999), as well as primates (Tokuda et al. 2002) such as squirrel monkeys (Brown et al. 2003) and baboons (Fischer et al. 2002), and also in humans in laughter (Bachorowski & Owren 2001), in infant cries (Mende et al. 1990), in children (Robb & Saxman 1988), and singing (Neubauer et al. 2004). Non-linear phenomena typically occur when F0 exceeds F1 due to instability at the source, with more bifurcations such as F0 jumps, subharmonics etc., typically observed in male participants with F0-F1 crossovers. This may be attributed to the fact that males are less accustomed to producing crossovers in everyday speech and therefore find it more difficult to overcome unwanted instabilities at the source (Titze et al. 2008). Crossovers typically hinder perception of close, front vowels in female speakers, because the F1 is most likely be greater than F0 in this context, but they may also explain why some distress productions, whereby both males and females may produce a higher F0 than their F1, are perceived as lacking in intelligibility.

The two explanations are not mutually exclusive and appear to demonstrate that productions of high F0 are likely to be modified through 'push' effects (physiological, reflexive responses) rather than 'pull' factors (external factors such as social norms) (Scherer et al. 1980). It also lends support to the implication in Johnstone & Scherer (2000: 222) that the more extreme the display of emotional expression, the more it is to be modified by push effects. Johnstone & Scherer (2000) also argue that more extreme emotion is perceived as more sincere, though this is not borne out in the present experiment.

Turning our attention to another acoustic parameter showing significant variation, intensity did predictably increase as F0 increased for both actors and victims in the speech samples. However, the actors were quieter than the victims when performing distress across samples. This is surprising since, although an independence of F0 and intensity is possible in human speech prosody (Demolin 2007), the physiological response to (di)stress promotes increases in both F0 and intensity as a consequence of the increased tension in the vocal folds and increased sub- and supra-glottal pressure (caused by the increase in respiration rate and muscle tension). For the victims, the two parameters behave as predicted, but for the actors there is a complication. I propose that the actors' deviation from the standard pattern is a

consequence of their dramatic training and, in some cases, as a means to avoid playing a stereotypical distress victim. Firstly, minimising vocal effort, e.g. singing high-pitched songs an octave lower or speaking passages instead of singing them, helps to avoid long term vocal damage (Webb 2007: 30). Producing high frequencies is not only potentially harmful for the voice, it also takes a great deal of energy (Webb 2007: 30). For the actors in this study, the drama workshop in which they were recorded was advertised as a voluntary training opportunity. Some were employed in acting jobs at the time and so may have decided to reduce their energy expenditure during the workshop, in order to conserve their energy for performances for which they are paid. Reducing loudness is one way in which they could minimise vocal effort. Furthermore, high-pitched screams and vocalisations, which many of them performed, may have been produced with diminished loudness in order to reduce the possibility of vocal damage. Secondly, their avoidance of loud distress portrayals (relative to their non-distress speech samples) may reflect a conscious decision to minimise overacting, i.e. to stay away from stereotypical performances as seen in horror movies, in order to create impact via a 'serious' performance. To this end, it would be interesting to consider the actors' preconceptions of what constitutes a good and bad performance, as well as to assess the saliency of pitch and loudness to actors performing emotional material. Thirdly, the use of a controlled reading passage to elicit neutral, non-distress speech from the actors may have been interpreted as another performance task, with the reading of the control passage performed loudly and clearly, rather than being spoken normally in a way that is indicative of normal-level everyday speech.

F0 and intensity variability are two examples of the wide range of variation that occurs in acoustic correlates of distress among victims and actors. It shows that the human larynx, when responding to a life-threatening situation, is capable of producing sounds that are typically not found in everyday speech, and indeed may result in speech productions that are comparable to those produced by individuals who have been trained to produce extreme material, e.g. singers or actors. Non-verbal productions such as distress vocalisations and screams show similar characteristics to non-verbal behaviours in animals, and the use of F0 in emotional speech may lend further support to Ohala's cross-cultural and cross-species

'frequency code' hypothesis. In addition to extending our general phonetic knowledge of how the human larynx can change its functioning in a life-threatening situation, these findings are also of interest to the forensic speech practitioner.

Given that forensic casework may arise from a 999 call or a recording of a life-threatening situation, forensic practitioners need to be aware of the potential effects of such a situation on an individual's speech output. We are already familiar with some effects on speech in forensic situations, e.g. increases in F0 and F1 in telephone-recorded speech (see, e.g., Hirson et al. (1995), Künzel (2001), Byrne & Foulkes (2004), undershooting of F1 and overshooting of F2 in heroin speech (Papp 2008), or increase in F0 and task duration in speech affected by alcohol (e.g. Chin & Pisoni (1997), Hollien et al.(2001)). Similarly, practitioners should be aware that the acoustic parameters of emotional distress speech may pattern very differently from the speaker's every-day, non-distress speech (to which practitioners sometimes have access for comparison work). F0 in particular may exceed the individual's typical range to the extent that a distress scream or vocalisation may be misidentified as having been produced by a member of the opposite sex, or indeed not attributable to an adult but to a young child (Victim F) or even a cat (Victim G).

Moreover, the extreme values observed in F0 have further practical considerations. F0 is a parameter often investigated by forensic practitioners using speech software, typically Praat (Harrison 2004). Yet it is sometimes measured with difficulty if using Praat's in-built settings, especially if the F0 is high, such as when conducting acoustic analyses of children's speech (Khattab & Roberts 2010). Formant frequencies are particularly hard to identify and measure due to the widely-spaced harmonics of the high F0 (Huggins 1980), and formant measuring techniques have been subject to investigation in order to reduce errors in formant frequency estimation (e.g. Traunmüller & Eriksson (1997)). If analysing high F0 speech material spectrographically, caution should be exercised when using software-calculated F0 values, and values should be confirmed via auditory impressionistic testing as well as measuring manually directly from the waveform or from the harmonics of a narrowband spectrogram (or both of these). Changing the frequency display from 5 kHz to 8 kHz, as well as increasing the maximum formant track

display to 8kHz, is recommended as a first step (Khattab & Roberts 2010: 170). Increasing the pitch ceiling in Praat is also advised as Praat's default ceiling for calculating F0 is 600 Hz, and yet many of the speech productions of both male and female victims and actors in the present study exceed this frequency. Above all, awareness that distress speech may lead to atypical acoustic variation should be an element of the analyst's approach.

8.1.2 Exploring vocal correlates of distress

From the findings reported in Chapter 5, it was concluded that there were significant changes from normal speech to distress speech in the acoustic parameters of F0, intensity, AR, and vowel formant frequencies. Specifically, increases in F0 mean and range were observed, as well as decreases in intensity and AR Vowel formant frequencies also changed, typically in the direction of higher F1 in distress, though the formant data showed considerable variation. The increase in F0 and F1 was predicted, but the decrease in intensity and AR was unexpected. It is worthy of note that these parameters did not permit discrimination between distress in actors and victims. The fact that there no distinction was detected between productions of genuine and acted emotion lends support to a dimensional view of emotion (i.e. one that maps affective states onto a specific dimension or combination of dimensions), since it can be claimed that level of emotional activation (arousal) is responsible for acoustic changes, independent of the emotion under investigation (Johnstone & Scherer 2000:227). In this investigation, both groups produced emotionally-aroused speech (or something that sounded like it), yet only the victims were suffering from authentic distress.

Two principal reasons have been put forward in emotional speech research to account for why it is hard to demonstrate emotion-specific vocal correlates. Firstly, the majority of previous studies have focussed on only a handful of acoustic parameters. With the study of additional parameters, or different combinations of parameters, we will be able to more finely differentiate between the different types of emotion categories and the activation dimension. Previous studies tended to investigate simple acoustic parameters (such as F0 and intensity) as they are robust and simple to measure. However, it can be argued that such simple measures perhaps

do not encompass the finer nuances of emotion. Johnstone & Scherer (2000: 228) highlight that recent research looking at more complex parameters, e.g. formant analyses, spectral analysis, inverse filtering, etc., looks promising as a route towards finer differentiation between emotions. Secondly, as noted in §2.3, there has been a lack of standardised definitions across disciplines concerning affective states (Scherer 2005: 696). The same basic emotion may be represented by two or more terms, e.g. anger (rage versus irritation), and fear (terror versus anxiety) (Johnstone & Scherer 2000:228). In these examples, the terms represent extreme ends of the same scale of emotion. The concept of "families of emotion" has only recently been proposed by proponents of discrete emotion theories (see Ekman (1994)) but it may help tease apart emotion-specific acoustic correlates from general emotional activation (Scherer 2003: 233).

Until these two factors are addressed, it is indeed tempting to consider similar results from emotional speech studies as being related to levels of activation, rather than to the emotion per se. This does not mean that emotion-specific categories cannot be distinguished by vocal cues, but rather that they demonstrate how emotions with similar underlying activation states can be difficult to distinguish discretely. However, listeners are, on the whole, able to consistently differentiate different emotions with higher-than-chance levels of accuracy (Scherer 1989). The picture that emerges from this study is that acted distress and authentic distress may behave as though they are two separate emotions, since they can, to some extent, be distinguished perceptually, but are not well differentiated acoustically. However, this seems counter-intuitive, as some people might consider authentic and acted distress as variants of the same emotion. Therefore, an alternative treatment of acted and authentic distress data would be one that involves two variables, one being the displayed emotion (distress versus non-distress material), the other being authenticity (authentic versus acted material).[19] In order to distinguish these two, further investigations would be required as the current study does not tease apart these variables.

[19] It is worth acknowledging that actors portraying distress are unlikely to be experiencing anything resembling a true distress response. However, as there is no way for researchers to directly assess their mental state, the focus here is on categorising the surface manifestation of the emotion rather than the internal state of the actor.

8.2 Listeners' perceptions of distress

8.2.1 Perceptual differences between acted and authentic distress

The findings reported in Chapter 5 demonstrate that there is no behaviour with respect to any one acoustic parameter or combination of parameters that can distinguish the productions of real victims from those of actors. It concluded that forensic practitioners should therefore continue to refrain from, or at least exercise caution when, making psychological assessments of distress, at least on the basis of cues provided solely by these four acoustic parameters. However, results from the perceptual experiment in chapter 7 were encouraging in that some listener groups were able to distinguish authentic from acted portrayals of distress.

It was found that listeners having familiarity with distress data (police call takers and forensic practitioners) exhibited accuracy rates of between 57% and 63% (excluding 'no decision' responses). These rates are similar to those reported in emotion recognition studies. Scherer (1989), for example, estimates that an accuracy rate of approximately 60% is typically found in most studies of this nature, but it should be noted that instructions for the perceptual test run for the current investigation and those studies summarised by Scherer are different; most emotion recognition studies involve a closed-set choice from a certain number of emotions, whereas my listening test employed a non-numerical Likert scale and offered a 'no decision' option. For the closed-set emotion recognition studies, the rate of chance was much lower than in the present perceptual study. This suggests that although the accuracy rates in the current study are better than chance, they are not as good as the recognition rates reported in other studies. The difference between authentic and acted distress is potentially perceptible, at least for some listeners. However, the results of this study are complex, as our consideration of what should be classified as correct or incorrect can skew the statistical analyses. If we exclude 'no decision' responses and consider only correct and incorrect responses, we find that police call takers and forensic practitioners perform better than chance, but that lay people are performing at chance level. If, however, we deem a 'no decision' response to be incorrect, then we find all groups perform worse than chance. Furthermore, if we consider just the number of correct responses per group from a total of all responses, we find that no one group

performs statistically significantly better than another. Consequently, it can be asked whether the evidence for perceptible differences between acted and authentic distress, as well as whether the evidence that some listeners perform better than others, is compelling.

Given that in real-life, forensic practitioners have the option to refuse a case on grounds of it not being analysable, I have considered a 'no decision' score to be neither a correct nor an incorrect response. Forensic practitioners who choose to not make a decision are not necessarily making an incorrect judgment; they are instead judging the sample unsuitable for analysis. Owing to the experimental design, it is unknown which of the 'no decision' responses are the result of the listener being uncertain and which are due to the listener considering the material to be unsuitable for analysis. However, after the experiment, many forensic practitioners told me that they were often reluctant to make a decision based on such brief samples, i.e. they felt that material on which to make an informed decision was too scanty, and therefore they opted to not make a judgment. This is demonstrated by the fact that the forensic listeners had a significantly higher rate of 'no decision' responses than the other two groups. Moreover, I suspect that some forensic practitioners still forced themselves into making a decision in order to provide results that reflect their perceptions of real and fake distress. I was told post-experiment by some practitioners that had this not been a research exercise, but rather a genuine forensic case, they would have rejected more of the stimuli on the grounds that they would be unsuitable for analysis.

A possible explanation for the high frequency of 'no decision' responses is that it arises from the caution that rightly pervades the culture of forensic science. This perhaps could be counted in the favour of forensic practitioners, as it demonstrates awareness of the high stakes that might be involved in real cases. All practitioners must exercise caution when analysing material and arriving at a decision. The forensic practitioner, if in doubt, should refrain from giving evidence that might offer support for a certain conclusion. In real-life cases, the wrong decision may have grave consequences, e.g. the incorrect conviction (a false positive) or acquittal (false negative) of the defendant. The majority of forensic practitioners therefore adhere to

a decision-making process whereby forensic audio is first assessed for suitability and feasibility of analysis before any analytical work is conducted; if the material is considered unsuitable for analysis, it can be ruled out at this stage, rather than forcing an analysis that may lead to a questionable decision. Evidence for the forensic practitioners' cautious responses lies not only in their significantly higher rate of 'no decision' responses but also in their significantly lower reported confidence levels throughout the experiment. Furthermore, although not measured formally as part of the experiment, the time taken to complete the listening experiment was much longer for forensic practitioners than it was for the other two listener groups. Forensic practitioners took, on average, 40 minutes to complete the experiment, whereas the lay people and police call takers typically completed it within 25 minutes.[20] The forensic practitioners were not using this time to repeat playback of audio stimuli (which was expressly forbidden), but to reflect and think before marking their decision on the response sheet.

The results discussed in both Chapters 5 and 7 provide further evidence of the existence of the paradox that production studies often lack emotion-specific vocal cues but perceptual studies reveal accurate levels of emotion identification (Scherer 1986: 143). Such results complicate the current situation whereby IAFPA members are prohibited from judging psychological states (and sincerity) as part of the IAFPA code of practice (IAFPA 2004). On the one hand, we are unable to put forward a profile of acoustic correlates that distinguishes real from acted distress. On the other, we find that listeners who have familiarity with distress material are able to differentiate between acted and authentic emotion with a reasonable degree of success.

Being able to distinguish between authentic and acted distress is of practical relevance to forensic practitioners (and potentially to emergency service personnel as well) since it is a task they are occasionally asked to perform. Police officers may

[20] The police call takers took part in the experiment as an authorised break during their shift, assuming that call frequency was low. Although some were grateful to not be in the call centre during this time and were therefore happy to spend longer on the experiment, others were eager to get back to their work and rushed through the experiment.

ask forensic practitioners whether audio recordings purporting to represent violent events are real or hoaxes, e.g. faked kidnappings. They may also question whether, and to what extent, vocalisations occurring in recordings reflect real distress, e.g. if sounds are consistent with one person hitting another, and whether the subsequent vocalisation from the victim reflects authentic pain and distress, or whether it was produced as an afterthought to make the assault more incriminating (mainly relevant if those involved are aware that their actions are being recorded).

8.2.2 Exploring listeners' accuracy

The findings of the listening experiment reported in Chapter 7 revealed that police call takers and forensic practitioners performed better than the lay people, but in different ways. Forensic practitioners had the lowest incorrect response rate but the greatest number of 'no decision' responses, whereas police call takers had fewer 'no decision' responses but a rate of correct responses similar to that of the forensic practitioners. Both groups have varying degrees of exposure to forensic material, and the forensic practitioners also have advanced phonetic/acoustic training and listening skills. Although it is difficult to quantify the role these two factors play, it appears that accuracy improves as a function of at least one of them. The table below highlights listener characteristics that may improve the listeners' ability to distinguish between authentic and acted distress. Natural aptitude is listed, since in all groups there were some listeners who performed better than others (as is expected just as a function of human variability).

Table 8-3: Listener characteristics that may influence ability to distinguish acted from authentic distress.

	Lay people	Police call takers	Forensic practitioners	Phoneticians
Natural aptitude	+/-	+/-	+/-	+/-
Exposure to real data	-	+	+	-
Advanced phonetic listening and training	-	-	+	+

Although there are several ways in which the experiment could be extended (see §9.2), one way in which the relationship between exposure to real data and advanced

phonetic training could be further investigated would be to conduct the same experiment on a new population, a group of phoneticians with no forensic training or experience (marked in Table 8-3 using dotted lines), to explore how their responses pattern compared to the other listener groups.

The results of the listening experiment raise some interesting further questions. Firstly, it is anticipated that the preconceptions or expectations of acted and authentic distress are different among lay people and those with familiarity with distress data (i.e. the forensic practitioners and the police call takers), since the lay listner's exemplars are likely to be based on media exposure and not the every-day analytical work performed by police call takers and forensic practitioners. An open question is "should there be encouragement to change exemplars of distress for lay people?" In addition, this has impact on drama professionals such as actors and directors. Whether total realism is good art is an enormous question (Morwenna Rowe, p.c.). Should drama professionals aim to be naturalistic when portraying emotion, and perhaps therefore risk an extreme emotion performance being misidentified by the audience, as shown by the responses to Extracts D and AM in the perceptual experiment, in which some victims giving genuine distress responses were misidentified as actors? Or should their intention be to act in a way that is recognised by the audience, but risk their performance being considered stereotypical or even bad, as illustrated in the responses to extracts AP and AA?

Secondly, a correlation was found between listeners' level of confidence and accuracy rate, albeit in a different direction from that which might have been predicted. In the literature concerning earwitness performance, studies investigating the relationship between voice identification performance and listener confidence have either found a positive correlation, where confidence is a reliable predictor of accuracy (e.g. Clifford et al. (1980)), or they found a non-significant (or weak) correlation between confidence and accuracy (e.g. Yarmey (2001)). In the current perceptual experiment, there was a negative correlation between level of confidence and accuracy rate. For police call takers and lay people, the more confident they were, the less likely were they to categorise the extract correctly. The forensic practitioners, however, did not demonstrate the same correlation. Given these

findings, those who perform an analysis in which they assess authentic and acted distress should practise extra caution in their decisions, regardless of their confidence level.

Thirdly, for emergency call takers, hoax calls are a perennial problem, as they divert emergency services from those who genuinely need them. For example, from April 2011 to March 2012, the Yorkshire Ambulance Service received over 750,000 urgent and emergency calls, of which 2,274 were hoax calls (Communications Department, Yorkshire Air Ambulance, 2012). Since call takers are trained to not question the integrity of callers, each call is treated as genuine, and it may lead to an emergency response. From the perspective of the police call takers, the listening experiment had no immediate practical benefit. The findings of this perceptual experiment do not change or improve the procedures that they already adopt. However, anecdotally speaking, the police call takers reported that if they do doubt the integrity of a speaker, they are able to collect additional data by asking more questions and listening for inconsistencies. This is very unlike the task performed by forensic practitioners, who often do not have access to, or the means to collect, additional data when analysing material for a report. They discount content and look for evidence in the form of the speech spoken, and they do not need act there and then. They have the ability to be slow and careful in their analyses. If any doubt remains, then the police call taker is expected to treat the call as genuine and will typically send an emergency response unit to further assess and handle the situation.

Finally, in the field of forensic speech science, if an opinion is sought about real versus acted distress, the forensic practitioner may be the best person to judge, given their higher accuracy rate (in this experiment) coupled with their caution and lack of hesitancy to reject unsuitable material. As mentioned in the previous section, it remains debatable whether they perform well enough to consider relaxing clause 9 of the IAFPA code of practice (the prohibition of IAFPA members to conduct psychological assessments of speakers). Ultimately, this question could only be resolved by those with executive responsibilities within the professional body of the association. An additional question arising from the findings is whether forensic practitioners might benefit from working alongside police call takers in order to

reduce their number of 'no decision' responses. Should an instructing party want a trained analyst to report on whether recorded distress was real or fake, s/he may find that a forensic practitioner rejects the recording and the question will remain unanswered. However, if s/he asks a police call taker to make a judgment, the police call taker may be more likely to get it wrong. This neatly parallels the debate within the Language Analysis for the Determination of Origin (LADO) field, in which practitioners have debated the merits of using native speakers and/or trained phoneticians to make decisions concerning an asylum seeker's origin (see Fraser (2009), Cambier-Langeveld (2010), and Nolan (2012)). In both cases, trained listeners may be assisted by individuals with greater previous exposure to, and familiarity with, the data with which they are working.

An interesting question arises from undertaking this particular task, which appears applicable to other areas across forensic speech science: what is (or should be) the direction of knowledge-sharing that can optimise our efficiency in our forensic endeavours?

8.3 Chapter summary

This chapter reviewed the findings from the investigation in terms of the two main research questions first presented in Chapter 1 of the thesis. First, it highlighted the special case of F0 and intensity as acoustic cues to distress, and discussed extreme vocal changes in speech when the speaker is under distress. Considerations for those analysing emotional speech were put forward. The second section showed that only some of the predictions about which acoustic parameters would distinguish distress from non-distress speech were borne out. A further complication in the form of the acoustic/perceptual paradox, in which the study of acoustic cues does not provide consistent data to put forward specific emotional vocal profiles, even though listeners are able to perceive this distinction, was discussed. Finally, listeners' performance when assessing acted and authentic distress was considered in relation to the IAFPA code of practice. Overall, the conclusion was reached that acoustic parameters probably only form part of the considerations that listeners take into account when identifying emotion in speech and other vocalisations.

9. Conclusion

This chapter summarises the key findings and implications of the research conducted for this doctoral thesis. It presents some brief ways in which distress speech research can be expanded.

9.1 Contributions to the field

This thesis represents the first step towards the empirical study of speech in distress. It provides a unique opportunity to investigate speech production under extreme circumstances; no other study has been able to explore the capacity of the vocal mechanism when the speaker is experiencing a violent attack. It is the first to directly compare authentic distress material uttered by real-life victims with re-enacted material performed by actors. Acted and authentic distress are investigated not only acoustically, but also perceptually. This innovative methodology enables the comparison and contrast of naturalistic distress vocalisations with acted portrayals of distress and can be extended to other production and perception studies of emotional speech.

At the heart of the thesis were two principal research questions. First, to what extent can specific acoustic measures be used to identify distress speech? The approach to answering this question included characterising distress speech, using acoustic features to distinguish authentic productions of distress from acted ones, and delimiting the boundaries of the individual's vocal performance. Secondly, can listeners perceive the difference between authentic and acted distress? If so, realted sub-questions can be raised: does listeners' accuracy and confidence vary as a function of familiarity with authentic distress material, and can listeners distinguish male and female distressed voices? To attempt to answer these questions, a perceptual experiment was conducted using extracts of authentic and acted distress, and it compared the performance of lay listeners, police call takers and forensic practitioners.

9.1.1 The acoustic study of distress

Chapters 3, 4 and 5 concerned the acoustic study of distress. Chapter 3 presented the methodology. It introduced the acted and authentic datasets and the acoustic parameters to be investigated: F0, intensity, AR and vowel formant frequencies. Analysis techniques were described and a taxonomy of distress productions introduced. Chapter 4 reported on a perceptual experiment that was conducted in order to test the reliability and replicability of the taxonomy prior to its use in the acoustic analysis. A secondary aim of the experiment was to test the influence of context on listeners' perceptions of distress. Listeners were asked to categorise extracts of distress productions based on the existing taxonomy. Some extracts contained the distress production in isolation, while others contained longer extracts with additional semantic material included in the extract. The findings showed that the taxonomy was reliable to some extent, and some modifications to the taxonomy were proposed and implemented before the acoustic analysis took place. It was revealed that both context and forensic experience led to significantly different perceptions of distress. Chapter 5 recounted the results of the acoustic analysis. It concluded that acoustic parameters can be used to distinguish between reference and distress conditions for actors and victims, but were not so helpful in discriminating between actors and victims. Specifically, it illustrated that the mean and standard deviation of F0 increased in distress speech produced by both actors and victims, that intensity decreased in the actors' distress speech, that the AR decreased in distress speech in both actors and victims, and that vowel formant values were subject to change when produced in distress speech by both actors and victims, but a systematic pattern to this change was not observed.

The major findings and contributions revealed in chapters 3-5 can be summarised as follows:

- A taxonomy of distress productions has been proposed and introduced as a way to categorise distress material;
- Naturalistic distress has been directly compared and contrasted with acted portrayals of distress, using the original emergency services transcripts as a basis for scripts with which to provide the actors;

- It was revealed that some acoustic parameters can be used to distinguish between reference and distress conditions, but are not helpful as discriminating features between actors and victims;
- A wide range of variation was observed across individuals' distress responses, demonstrating that there is no one 'real' portrayal of distress;
- Given that acoustic cues could not be used to differentiate between authentic and acted distress, it added justification to the current FSS practice of refraining from conducting psychological assessments in forensic casework;
- It was observed that reference material follows the expected pattern of acoustic parameters, whereas distress speech deviated from reference material (though not always in the same way across individuals), and so it may be easier to demonstrate which acoustic properties are *not* usually found within distress material, e.g. it typically does not exhibit a stable mean and standard deviation in either F0 or intensity, it does not typically fall within the reported AR ranges for speech, and it may/may not produce typical vowel formant frequencies;
- The extreme nature of vocal productions of distress, as exemplified by some hitherto unreported F0 values of individuals screaming and the presence of non-linear phenomena, was illustrated, extending our knowledge of the limits of vocal performance.

9.1.2 The perceptual study of distress

Chapters 6 and 7 concerned a perceptual study to investigate listeners' perceptions of acted and authentic distress. Chapter 6 introduced the experiment methodology. It described the selection of audio stimuli extracted from the recordings of authentic and acted distress used in the acoustic study. A description of the participant selection and recruitment was also described. The listener groups were comprised of lay people, police call takers, and forensic practitioners. The design and procedure of the experiment was also reported. Chapter 7 described the results of the perceptual experiment. It revealed that listeners with experience of authentic data (i.e. forensic practitioners and police call takers) were able to differentiate between acted and authentic distress, but that lay listeners performed no better than chance level. The forensic practitioners were the least confident listener group, though all groups

tended to express low levels of confidence. All listeners performed well in correctly identifying the sex of the speaker in each extract, though some extracts produced by a male with a high F0 were misidentified. No significant differences between listener groups were observed in this latter task.

For the perceptual study of distress, the major findings and contributions revealed in chapters 6 and 7 can be summarised as follows:

- Acted and authentic extracts of distress speech were used as stimuli in a listening experiment, in which three groups of listeners - lay people, forensic practitioners and police call takers - were invited to rate the extracts for accuracy, confidence level of the listener, and sex attribution;
- Listeners performed no better than chance at distinguishing between authentic and acted emotion, but police call takers and forensic practitioners achieved accuracy rates greater than chance;
- Police call takers attained a high number of 'no decision' responses and gave the highest rate of correct responses, whereas forensic practitioners gave the greatest number of 'no decision' responses and the lowest number of incorrect responses, which indicates that both listener groups have a skill that helps them to perform better than lay listeners at the this task. The experiment design did not, however, allow this to be investigated further;
- The role of both police call takers and forensic practitioners could be useful in the assessment of distress since both groups have some skill to offer when making a judgment about acted vs. authentic distress;
- The difference in performance by police call takers and by forensic practitioners parallels neatly with the performance of native speakers and trained linguists in LADO case work. The debate concerns disagreements with respect to who is better equipped to perform the task of determining distress authenticity/the origin of a speaker. In the perceptual study of distress, both forensic practioner and police call taker listener groups perform better than chance but in different ways. The forensic practitioners' caution reduces their rate of misidentifications, but increases their number of 'no decision' responses. On the other hand, the police call takers attempt

more judgments, but have a greater number of incorrect responses. Ideally, a low rate of incorrect responses coupled with a low rate of no judgments would be the most desired outcome;

- Forensic practitioners were the least confident listener group, though low levels of confidence were obtained amongst the majority of participants;
- For many extracts, listeners were able to identify the sex of the speaker, yet some exceptions were noted whereby the F0 was atypical of the range expected of that sex.

Taking into account both the acoustic and the perceptual study, it can be observed that an acoustic profile of distress would be difficult to propose, given the variability in the data, but some generalisations can be made. This observation nonetheless lends support to the IAFPA ruling in the code of conduct that prohibits forensic practitioners from assessing emotion or sincerity in forensic casework. However, given the results of the perceptual experiment, in which listeners who have familiarity with forensic material appear better equipped to recognise genuine distress than those who have not, the role of the forensic practitioners in assessments of emotion should perhaps still be further explored before any conclusion can be drawn.

The discussion presented in Chapter 8 reviewed the findings from both the acoustic and perceptual studies. First, it considered the special properties of F0 and intensity in distress productions, and highlighted some extreme vocal changes, such as non-linear phenomena, that occur when producing distress responses. It also explored potential reasons why distress responses have evolved to have these features. In light of this, some considerations for those conducting analyses on emotional speech in general, and distressed speech in particular, were put forward. Next, it discussed the apparent paradox arising from the fact that listeners are able to distinguish speaker-specific emotions in perceptual studies even though the production studies have not yet found any unambiguous and consistent acoustic cues for distress. Finally, the differences in performance between the three types of listeners were considered, and explanations for the different strategies they followed while judging acted and authentic extracts were proposed. Given that no group's performance was

without considerable error, it was argued that these studies support the current IAFPA code of practice, with the caveat that if an authenticity judgment is deemed absolutely necessary, forensic practitioners may be the best people to perform it.

9.2 Future research

Although the current investigation acts as a modest first step into distress speech research for forensic purposes, the research can be extended in a number of ways. Firstly, the perceptual experiment can be expanded to involve new listener groups and longer extracts. Adding a group of non-forensic phoneticians as a listener group would help to investigate whether phonetics training versus familiarity with authentic data contribute to improving listener accuracy. Similarly, it would be beneficial to introduce a group of drama professionals as a listener group to explore how their perceptions of distress differ from those of the other listener groups.

In the current study, many participants reported that the extract was too short to allow them to make a decision. The experiment could be expanded to include longer extracts so as to investigate whether additional material assists the listeners when making their decisions about the acted or authentic distress extract in terms of the resultant accuracy rate and level of confidence.

The presence of non-linear phenomena in the spectrographic examinations of the data was unexpected. More research could be conducted in this area to highlight which conditions are most likely to result in bifurcations in the data. The effect of non-linear phenomena on listeners' perceptions of distress could also be investigated via another listening experiment.

Finally, distress screams have recently featured in high-profile legal cases. In the *State of Florida -v- George Zimmerman* trial, the question was whether screams could be attributed to an individual based on his 'normal' (or reference) speech. During the early stages of *The State vs. Oscar Pistorius* trial, the attribution of a high-pitched scream to a man or woman was under debate. There is little research on screams, and yet they are clearly of legal interest in some criminal cases. As an extension of the current study, I intend to collect a corpus of screamed and reference data from actors in order to conduct a listening experiment to investigate whether familiarity with an individual's 'normal' voice, i.e. reference speech material,

enables listeners to identify that individual through his/her screams. Furthermore, by including a perceptual experiment that includes both male and female reference voices, the ability of listeners to correctly identify the sex of the screamer could be further explored.

The research in this thesis, therefore, opens up paths of investigation into the analysis of distress and of screams in forensic contexts that have not been considered earlier. Regardless of what will actually be found in these further investigations, it is clear that these data are worthy of empirical investigation, and that this thesis can serve as a foundation for this research moving forward.

Appendix A - Transcripts

Transcript A

Background: 2 male colleagues, Bill and Steve, are in the front of a car. They offer a lift to a male hitch-hiker on their way home. Everything is going fine until the hitch-hiker complains of feeling unwell. They pull over and both Bill and Steve get out the car to check on him. Steve sees through the window that the hitch-hiker has a gun. As he turns to run away, shots are fired. Steve is shot but manages to find cover. Bill is lying injured near the car. The hitch-hiker still has the gun and as he walks towards Bill, Bill pleads for his life.

N.B. All names and personal information, including location and telephone numbers, have been changed.

Conventions of Transcript:

B	Bill
H	Hitch-hiker
-	Hyphens denote incomplete or interrupted speech
[coughing]	Description of non-verbal sounds

[Start]

B: So, what happened to your car?

H: I don’t know. It broke down. It was going to take ages for them to come pick it up so I thought I’d try to get a lift into the town and sort it out from there instead.

B: At least you didn’t have to wait around in the rain until we offered you a lift.

H: Yeah. How long is this ride going to be? I just want to get there.

B: Don’t worry, we’ll be there soon. Just try and relax for a bit, don't get yourself all wound up, eh!

H: I know, I mean, I also have to take some medication.

B: What medication are you on?

H: I need to take some insulin every day.

B: What's that for?

H: Metabolism disorder – I’m diabetic.

H: I'll be gone all day. When do we get to the city?

B: Don't know yet, mate. Depends how long it takes in the traffic.

[Whistling, humming, tapping]

B: [directed to Steve] Hmm, should I take the main road in? Take the main road in? Avoid that country lane and go in via the North?

[Steve shrugs]

B: Up to you, do you want to? Be safer, anyway.

[Steve nods OK]

H: Woah, I got to…

B: You alright?

H: Can you open the windows?

B: I'll pull over.

[Car pulls over, Bill and Steve get out of the car to check on the hitch-hiker. Steve sees the gun and is able to say "He's got a-" before shots are fired]

B: Argh! Argh! Argh!

(S runs for cover but B has been injured and can't run. H walks toward B)

B: Oh. No. Don't you- Please. Don't.

[End]

Transcript B

Background: A known drugdealer, Phil, is conducting a deal outdoors one evening but it goes wrong when his buyers want more. He is stabbed in the stomach by one of the buyers. They run off and he is now alone. He calls 999 and speaks to the operator. During the conversation, the buyers come back in a car.

N.B. All names and personal information, including location and telephone numbers, have been changed.

Conventions of Transcript:

AS	Ambulance Service
P	Phil
O	Operator
-	Hyphens denote incomplete or interrupted speech
[coughing]	Description of non-verbal sounds

[Start]

AS:	Ambulance Service.
P:	I've been stabbed mate, I've-
AS:	Just a moment.
P:	Operator, I need help. I've been stabbed.
AS:	Operator-. Where, just a moment- operator, give me the telephone number, please.
O:	It is a mobile number 01234-
P:	Yeah, yeah
O:	01234 567891
P:	Come on, I'm bleeding like mad here.
AS:	Just a second, sir. Go ahead operator.
O:	01234
AS:	Yeah
O:	567
AS:	Yeah

Operator:	891
AS:	That's lovely, thank you. Where are you? What's the address?
P:	I'm on the, I'm down on the, I'm down by the Cricket Mill Plain.
AS:	You're down by where?
P:	Cricket Mill Plain, down by Spokesly. Quick, I'm bleeding. I've been stabbed in the stomach.
AS:	You've been stabbed in the stomach?
P:	Yeah, Yeah.
AS:	You're by Cricket Mill Plain?
P:	Yeah.
AS:	Where is that please?
P:	Spokesly, down by Spokesly, right by Goodside. Come on, I think that I'm going to pass out, very quickly. Honestly, I'm bleeding like mad here.
AS:	And it's Spokesly you're saying?
P:	Yeah, like, I'm on the road, on the main road, going up by the, by the Bay Horse turning.
AS:	Hello?
AS:	Hello?
AS:	Hello?
AS:	Hello?
[vehicle audible]	
Phil:	Argh. Argh. Ah, Jesus.
[vehicle audible]	
[End]	

Transcript C

Background: A landlord, Hugh, hasn't received the rent from his lodger, Chris, and so has confronted him. Chris turns nasty and picks up a knife from the kitchen, stabbing Hugh in the neck. He runs off leaving Hugh alone in the house. It is not known whether Chris comes back during the course of the conversation.

N.B. All names and personal information, including location and telephone numbers, have been changed.

Conventions of Transcript:

AS	Ambulance Service
H	Hugh
-	Hyphens denote incomplete or interrupted speech
[coughing]	Description of non-verbal sounds

[Start]

Ambulance Service (AS):	Ambulance Service. Can I help you?
H:	Please. Hurry Up.
AS:	Hang on
H:	Ambulance. [inaudible] -close. Emergency. I'm-
AS:	Yeah. Alright. Calm
H:	-can't. I'm bleeding to death. I've been stabbed through the th- …
AS	What's the address? What is the address?
H:	Six hundred and nineteen Pages Road, Nins-wood. Six hundred and nineteen Pages Road, Nins-wood. …- pumping blood out. I'm pumping- [screaming] blood out. Argh! Argh! Argh! Argh!
AS:	Is it for yourself?
H:	Yeah.
AS:	Is it for yourself?
H:	Yeah. I'm dead.

[Rustling]

AS: Hello, Operator?

[End]

Transcript D

Background: A teenage girl, Rose, has accused her uncle, Jason (Jase for short), of looking at indecent images on a family computer. She confronts him with her twin brother. The uncle later asks to meet them both to talk about it. They go to his house but during the conversation he becomes threatening and produces a loaded gun. They run upstairs but he is able to shoot Rose's brother in the head. Rose locks herself in the bathroom and calls 999 from her mobile. Jase manages to kick down the door and shoot Rose.

N.B. All names and personal information, including location and telephone numbers, have been changed.

Conventions of Transcript:

R	Rose
PS	Police Service
O	Operator
-	Hyphens denote incomplete or interrupted speech
[coughing]	Description of non-verbal sounds

[Start]

O:	I'm connecting mobile 01234 567891
R:	[screaming] Jase, let go of me .... Jase ... Jase
PS:	Thank you, go ahead, Caller. How can I help?
R:	[screaming] Jase don't. I'm telling you-
PS:	Hello, Caller?
R:	Let go of me.
PS:	Hello?
R:	Get off me.
PS:	Hello, what's your address?
R:	Help.
PS:	What's your address? Tell me your address?
R:	[screaming] Jase.

PS: Hello?

R: Jase.

PS: Hello?

R: Stop it. Jase. What are you doing?

PS: Hello?

R: Jase. [pause] Get out.

PS: Hello?

R: Get out, Jase. Leave me alone.

PS: Hello?

R: Help [Heavy breathing]

PS: Hello?

R: …uncle's just killed my brother. [pause] We were-, he was-, getting bad tempered-, there was no point-

PS: What's your address?

R: Eight three four Boston Close, Mast-in-Surrey.

PS: Ha- hang on. Eight-

R: He's still got a gun in his hand

PS: Hang on, slow down. I can't tell what you're saying. Tell me.

R: My uncle's just killed my brother. I-, we were-, he went made-, he's still mad-, I've got no choice-

PS: Alright, now tell me your address.

R: Eight three four Boston Close

PS: Eight three four Boston Close in where?

R: Mast-in-Surrey

PS: In where?

R: Mast-in-Surrey

PS: In Mast-in-Surrey alright, stay on the phone with me.

R: Help

PS: Alright you stay on the phone with me and tell me what's happening till we get somebody to you, alright.

R: Yeah [heavy breathing]

PS: Okay, what's your name?

R: Rose.

PS: Alright, Rose

R: Help.

PS: Alright we're going to get somebody to you right now while I'm on the phone to you, alright?

R: No I-can't stay here.

PS: What's your surname ,Rose

R: Eh?

PS: What's your surname ?

R: Rogerson. I've locked myself in the bathroom.

PS: You are, are you?

[rattling]

R: [Screaming] No, no, no, no, Jase, no, don't. Argh. Argh. Argh.

[shot]

[moaning]

PS: Hello? Hello, Rose? Hello, Rose? Rose? Rose?

[moaning]

PS: Hello, Rose? Rose, can you hear me?

[moaning]

PS: Rose, can you hear me? If you can hear me, tap something for me.

PS: Rose, if you can hear me tap the phone for me.

PS: Rose, can you hear me? Rose, if you can hear me can you talk to me?

PS: Rose I need you to try and let me know you can hear me? Can you hear me?

[End]

Transcript E

Background: A young woman, Jane, is returning home from work early. As she walks through her front door, she startles her housemate who is showing off a pellet gun to some friends who are round. Jane is accidentally shot through the chest. Her housemate is frozen in shock but Jane is able to call 999.

N.B. All names and personal information, including location and telephone numbers, have been changed.

Conventions of Transcript:

J	Jane
O	Operator
AS	Ambulance Service
-	Hyphens denote incomplete or interrupted speech
[coughing]	Description of non-verbal sounds

[Start]

O:	Emergency, which service?
J:	Hello?
O:	Do you need fire, police or ambulance?
J:	An ambulance, please. My chest feels tight and I don't like it.
O:	One moment, please. I'll put you through to the ambulance service?
J:	[addressing housemate] Stay here, won't you, until some comes?

[pause]

O:	Sorry to keep you waiting, I am trying to connect you.
AS:	Ambulance control?
O:	London connecting 01234 567891
J:	Pardon? I don't know what you're saying.
AS:	Thank you. Where do you want the ambulance sent?

J: I’m at three one four Oaks-Vale Street. I’m scared.

AS: Oaks-Vale Street. OK, what’s the problem?

J: I’ve been shot.

AS: You’ve been shot, have you?

J: Been shot.

AS: By what? By who?

J: By a gun.

AS: What a pellet gun or a proper gun?

J: Come on…

AS: Sorry? Talk to me.

J: …

AS: Hello?

[End]

Transcript F

Background: A successful businessman, Will, is duped into letting a con-man enter his home. On coming in the house, the conman shoots Will in the back and then shoots his wife, Marie, through the head/neck and proceeds to burgle the property. Although severely injured, Marie crawls to the telephone and calls 999 while the conman is still inside the house. Due to her injuries, Marie has difficulty being understood by the operator. The conman hears her attempts and goes to shoot her again.

N.B. All names and personal information, including location and telephone numbers, have been changed.

Conventions of Transcript:

M	Marie
O	Police Service
-	Hyphens denote incomplete or interrupted speech
[coughing]	Description of non-verbal sounds

[Start]

O: Emergency, which service?

M: Police. Please.

O: Pardon?

M: Please. Police.

O: Is mummy there?

M: [vocalisation]

O: Is mummy there?

[audible breathing]

[footsteps]

[vocalisation/scream]

M: No. Argh.

[End]

Appendix B - Workshop Schedule

The workshop takes place at **Jerwood Space**, Union Street in London starting from **10am** and is due to finish by **6pm.** Information concerning getting to Jerwood can be found on the next page, after the schedule.

We will break for lunch around 1pm but it's likely that few places will be open nearby given that it's a Sunday. You may want to bring your lunch with you. Also, remember to bring plenty of water with you - you'll need it throughout the day, especially during voicework.

Extreme Emotion Workshop - July 4th

10.00 – 10.15	Introduction to workshop (Lisa, Morwenna, Dan)
10.15 – 11.00	Physical and vocal warm ups (Morwenna & Dan) Preliminary recording (Lisa)
11.00 – 1.00	Session I: 'outside to in' skills-based techniques (Morwenna)
1.00 – 2.00	Lunch Individual recordings (Lisa)
2.00 - 4.00	Session II: 'inside to out' psychological acting techniques (Dan)
4.00 - 5.00	Break Individual recordings (Lisa)
5.00 - 6.00	Session III: Extreme emotion in forensic cases (Lisa)

If you have any queries or concerns about the day, please don't hesitate to contact Lisa on lsr501@york.ac.uk

Appendix C - Workshop Equipment List

Audio - Head mic

- DPA 4066 Head-band mic - with wind shields
- Marantz PMD 670 with power supply + compact flash card 255Mb
- Compact flash card 4 GB (spare)
- Compact flash card USB reader
- Mic-XLR adaptor
- XLR cable
- Mic stand

Audio - Zoom

- Zoom H4 with 128MB S.D. card and USB cable
- 2GB S.D. card
- Power supply

Video

- Panasonic S.D.R-H90 video camcorder - USB cable
- Camera tripod
- Camera tripod base for camcorder

Other

- Extension cable
- Beyerdynamic DT 250 closed cup headphones
- Sound level meter, incl. calibrator
- Tape measure
- Speakers

Appendix D - Equipment List

D1 Equipment used during the acoustic analyses

Hardware

- 2.4 GHz Intel-based PC with M-Audio sound card
- 1.73 GHz Intel-based laptop with Sigma Tel High Definition Audio CODEC
- Sennheiser HD 280 headphones

Software

- Sony Sound Forge (version 9.0c)
- Praat (version 5.0.22 and 5.1.25)
- Microsoft Office Excel (97-2003)

D.2 Equipment used in the re-recording of acted extracts

Hardware

- Motorola V500 mobile telephone, released 2003
- Nokia 2310 mobile telephone, released 2006
- Audioline (Business Class) AUB1landline telephone
- Prospect TC30 Telephone Interface Adapter (handset mic muted)
- Rane RS1 230 Vac remote power supply
- Rane MS-1 Microphone Preamp
- Sennheiser HD 280 Pro headphones
- M-Audio Delta Series sound card (break out box) – (Professional 4-In/4-Out Audio Card)
- Trust Soundforce 2.0 Speaker Set SP-2200 – PC multimedia speakers (single driver, wide bandwidth)
- 2.4 GHz Intel based PC with M-Audio sound card
- 1.73 GHz Intel based laptop with Sigma Tel High Definition Audio CODEC
- Stagg microphone stand - Black with Telescopic Boom Arm
- Clamp for microphone stand

Software

- Sony Sound Forge (version 9.0c)
- Praat (version 5.1.25)

Order of recording:

- Living room, flat – Case C
    - TB47M – Nokia

    - MS27M – Motorola
- Bedroom, flat – Case F
    - LE26F – Motorola
    - SS28F – Motorola
    - TM38F - Nokia
    - TM38F – control – Nokia
- Hallway, flat – Case E
    - SS25F – Nokia
    - DM27F – Motorola
    - DM27F – control – Motorola
- Bathroom, flat – Case D
    - ZC30F – Nokia
    - ZR28F – Nokia

    (no Motorola reception in the bathroom)
- Outside car park – Case A and B
    - RG27M – B42M – Nokia (x 2)
    - PB23M - A34M – Nokia
    - PW28M – B42M – Motorola
    - JS23M – A34M - Motorola

Appendix E - Taxonomy Experiment (Chapter 4)

E1 Information Sheet

Background

My PhD research investigates phonetic cues of distress using forensic audio data of people who have been subjected to extreme physical and emotional distress, e.g. a violent assault.

Binary distinctions such as 'modal' (i.e. 'everyday', 'non-emotional speech') or 'distress' do not take into account all the speaker variations that are often exhibited throughout the recording. Often, victims' speech appears to respond to different degrees of stress throughout an attack – presumably due to the different physical and situational stimuli - and vocal responses form a continuum of conditioned speech ranging from modal to extreme distress.

I have been using the following four-way scalar categorisation when coding the forensic data:

1. modal speech (non-emotionally aroused speech, intelligible)
2. distress speech (emotionally aroused, intelligible)
3. distress vocalisations (emotionally aroused, unintelligible)
4. distress screams (emotionally aroused, no linguistic content)

Experiment

The following experiment will involve listening to excerpts from authentic forensic material. Some excerpts will be played in isolation, for other excerpts you will be given some contextual information (both preceding and following phonetic contexts as well as background information about the case). You will hear each excerpt three times and will be asked to categorise the excerpt according to the labels given in the distress speech continuum (described above) on the response sheet provided. You will also be asked to rate each excerpt in terms of perceptible linguistic content. An empty notes section is included on the response sheet for each excerpt in case you want to make notes, IPA transcriptions, or highlight the excerpt for discussion following the experiment. Notes concerning characteristics of the speaker (e.g. male/female, old/young), the extent of distress perceived in the speaker, and your interpretation of what was said are particularly encouraged.

E2 Consent Form

By taking part in this experiment, you have been given access to data from real forensic cases. The materials contain incidents involving or referring to criminal activities, some of which may be considered disturbing. The recordings may also contain references to individuals' names and personal details. The following is a standard agreement for using these materials.

I accept the following conditions:

1) In participating in the experiment I acknowledge the sensitive and potentially distressing nature of the materials, and I confirm that I am willing to work with such materials.

2) I will not attempt to contact any individual whose speech or language forms part of the materials, or who is referred to in the materials.

3) I will not duplicate, circulate or otherwise transmit any materials used in the experiment (including sound files and PowerPoint slides) to any parties outside the context of the forensic research group.

4) It is understood that any work done using the materials is for the sole purpose of participating in the experiment. I agree to make no other use of the materials

(including publication and oral presentation of work) without entering into a separate written agreement with Lisa Roberts or Peter French.

5) After the experiment has ended I agree to destroy any copies of sound files and associated transcripts originating from the materials.

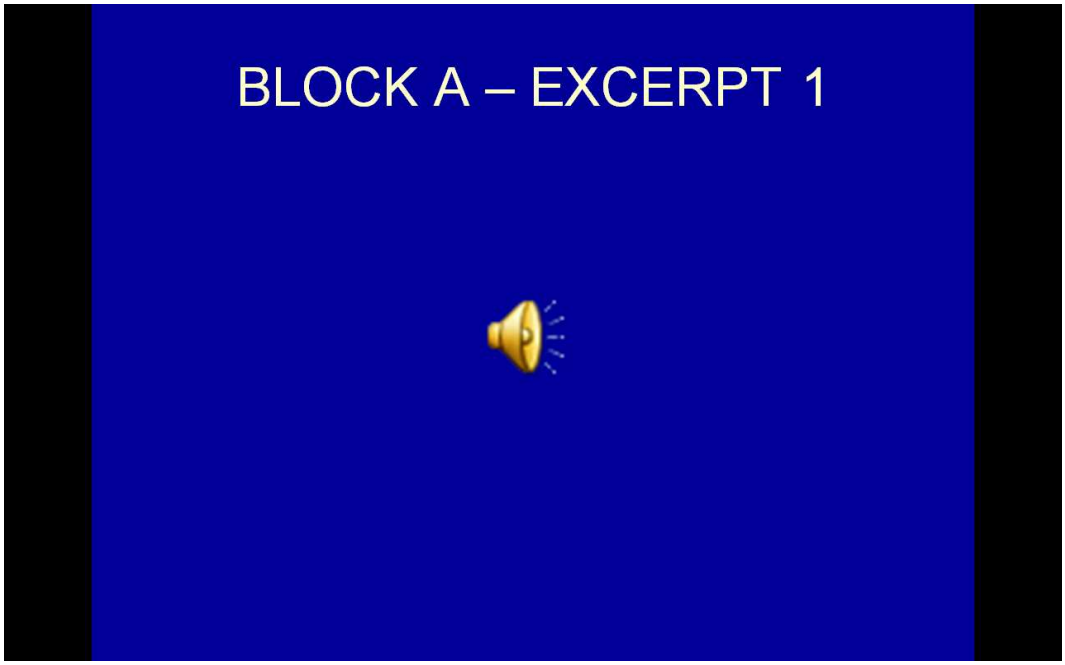
6) I understand I can withdraw my participation at any time, and that I am under no obligation to complete the experiment.

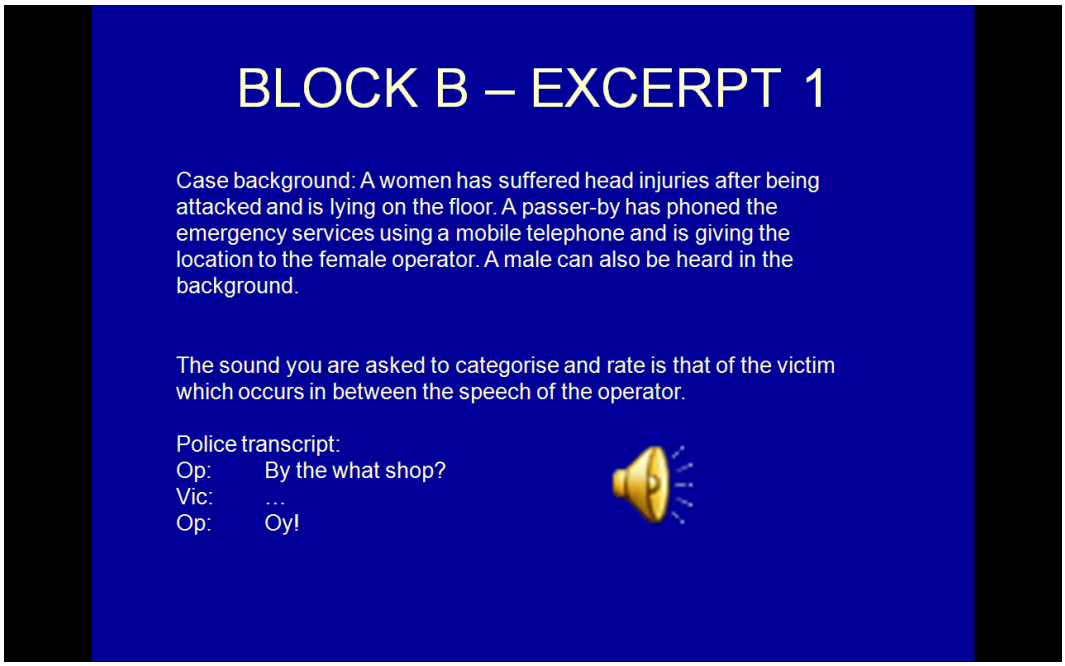
I have understood the conditions outlined in this agreement and agree to abide by them.

Name_____

Signature_____ Date_____

E3 Examples of visual stimuli of audio extracts with and without contextual information





E4 Response sheet

Experiment Info

(please circle)

Experiment: A B

Participant Info

(please circle or state where applicable)

Forensic Casework Experience: Experienced Inexperienced

Level of formal phonetic training: MA/MSc PhD PhD+

Sex: M F

Age: ……

Native Language(s) *(if not British English*) ……..
……..
……..

Block A Responses

Excerpt 1:

1. Would you categorise this excerpt as:

(please tick one box)

Modal	Distress Speech	Distress Vocalisation	Distress Scream

2. Please rate the excerpt in terms of perceptible linguistic content:

(where 1 = clear linguistic content, and 5 = no apparent linguistic content)

1	*2*	*3*	*4*	*5*

3. Notes:

Appendix F - Acoustic Findings (Chapter 5)

F1: Table showing F0 mean, standard deviation (S.D.) in Hertz (Hz), min., max., and S.D. in semitones (ST) for all actors across all conditions.

Speaker	Condition	mean (Hz)	S.D. (Hz)	min. (Hz)	max. (Hz)	S.D. (ST)
Act 1	ref	99.0	10.1	74.9	142.5	3.5
	un-r	233.6	92.8	85.2	584.1	14.6
	reh	238.2	62.2	117.9	455.3	9.2
Act 2	ref	97.0	17.0	74.0	196.0	6.1
	un-r	218.6	61.7	108.0	320.2	10.1
	reh	230.3	155.0	81.0	453.9	28.3
Act 3	ref	117.1	30.1	74.9	220.3	9.1
	un-r	187.0	40.6	83.4	360.8	7.6
	reh	204.0	70.5	84.6	443.6	12.5
Act 4	ref	95.0	8.5	75.6	135.8	3.1
	un-r	197.0	35.5	76.1	316.2	6.3
	reh	184.9	33.7	82.6	329.3	6.4
Act 5	ref	95.4	13.1	73.6	157.7	4.8
	un-r	246.3	95.2	112.3	509.9	14.1
	reh	257.8	60.8	134.8	393.1	8.3
Act 6	ref	88.7	13.5	74.0	154.9	5.3
	un-r	200.9	180.0	74.7	868.0	50.2
	reh	258.9	198.7	120.4	901.8	35.1
Act 7	ref	197.7	22.2	147.7	290.4	3.9
	un-r	x	x	x	x	x
	reh	505.3	153.8	90.2	1155.6	10.9
Act 8	ref	226.6	20.9	175.5	332.3	3.2
	un-r	438.4	98.2	173.6	723.6	7.9
	reh	451.6	128.0	182.8	865.7	10.1
Act 9	ref	200.9	36.5	115.1	341.6	6.4
	un-r	257.2	33.0	161.7	346.7	4.5
	reh	241.9	43.7	137.0	419.3	6.3
Act 10	ref	179.0	39.4	124.5	352.6	7.8
	un-r	383.1	124.5	198.0	804.8	11.7
	reh	377.7	63.9	247.3	612.5	5.9
Act 11	ref	173.0	22.1	77.9	290.6	4.4
	un-r	348.4	164.7	192.1	661.1	17.8
	reh	424.6	111.5	234.0	665.0	9.3
Act 12	ref	154.4	33.3	95.3	296.1	7.6
	un-r	183.5	33.9	127.9	259.0	6.5
	reh	765.7	181.7	75.1	1047.9	8.4

F2a: Mean intensity values (dB) for victims' distress responses as categorised by the distress taxonomy.

	Cat 2	Cat 3	Cat 4	Cat 3a	Cat 3b	Cat 3c	Cat 4a	Cat 4b	Cat 4c
Vic A	x	75.70	83.33	75.28	x	76.95	x	83.33	x
Vic B	66.12	65.66	x	65.91	x	65.16	x	x	x
Vic C	55.52	58.81	59.42	x	x	58.81	61.44	57.40	x
Vic D	74.69	74.70	74.42	74.70	x	x	74.22	74.25	75.01
Vic E	79.53	x	x	x	x	x	x	x	x
Vic F	x	71.99	69.54	73.19	x	71.39	x	x	69.54

F2b: Intensity standard deviation values (dB) for victims' distress responses as categorised by the distress taxonomy.

	Cat 2	Cat 3	Cat 4	Cat 3a	Cat 3b	Cat 3c	Cat 4a	Cat 4b	Cat 4c
Vic A	x	8.38	3.53	8.37	x	8.42	x	3.53	x
Vic B	12.32	8.83	x	7.62	x	11.26	x	x	x
Vic C	7.60	7.77	6.96	x	x	7.77	5.94	7.99	x
Vic D	5.15	6.22	3.63	6.22	x	x	3.58	4.02	3.54
Vic E	4.55	x	x	x	x	x	x	x	x
Vic F	x	9.59	6.57	9.89	x	9.44	x	x	6.57

Distress taxonomy categories:

1 = reference speech

2 = distress speech (intelligible, produced in a distressing, emotional context)

3 = other (unintelligible, produced in a distressing, emotional context)

- 3a - contains linguistic content
- 3b - does not contain linguistic content
- 3c - unclassifiable

4 = scream

- 4a - contains linguistic content
- 4b - does not contain linguistic content
- 4c - unclassifiable

F3: All formant values (Hz) for male victims in reference (where available) and distress speech.

Victim		Vowel	F1	F2	F3	From (s)	To (s)
A	distress	iː	480	1712	3529	8.59	8.6
	reference	iː	428	2114	2562	10.02	10.04
B	reference	iː	421	2165	3349	2.91	2.92
		iː	455	1829	2437	6.03	6.04
		iː	471	1907	2594	14.89	14.89
	distress	iː	451	2079	2935	9.11	9.12
		iː	397	2021	2612	17.21	17.23
	reference	ɪ	367	1670	2465	2.47	2.48
	distress	ɪ	432	1399	2464	5.03	5.03
		ɪ	458	1826	2221	6.41	6.42
		ɪ	473	2006	2723	6.6	6.6
		ɪ	415	1858	2180	8.26	8.27
		ɪ	483	1527	2209	15	15.01
		ɪ	402	1693	2240	16.21	16.22
	reference	ɛ	571	1739	2326	14.64	14.65
	reference	a	567	1345	2482	4.28	4.29
	distress	a	660	1413	2160	0.54	0.55
		a	576	1499	2276	9.64	9.65
		a	696	1289	2254	15.48	15.49
		a	662	1415	2282	17.61	17.62
	reference	ɑː	658	1049	2383	8.87	8.89
	reference	ɒ	512	1011	2038	1.84	1.85
		ɒ	655	1198	2251	1.03	1.05
		ɒ	700	1118	2254	0.35	0.36
	reference	ʌ	455	1107	2286	11.8	11.81
	distress	ʌ	462	1071	2303	10.15	10.17
		ʌ	622	1307	2311	21.18	21.19
C	distress	iː	310	1454	2469	2.1	2.1
	distress	ɪ	296	2092	2436	0.62	0.63
		ɪ	377	2308	2553	4.2	4.21
		ɪ	479	2362	2661	5.51	5.51
	distress	ɛ	534	1749	2512	14.05	14.07
		ɛ	591	1453	1766	2.37	2.38
	distress	a	840	1456	2312	2.82	2.83
		a	862	1381	2431	0.06	0.07
	distress	ʌ	644	1465	2543	6.39	6.4
		ʌ	719	1321	2447	6.99	7
		ʌ	737	1291	2110	5.99	5.99
		ʌ	767	1457	2407	7.41	7.42
	distress	u	645	1273	2423	5.76	5.76

F4: All formant values (Hz) for female victims in distress speech.

Victim		Vowel	F1	F2	F3	From (s)	To (s)
D	distress	iː	473	2129	3093	3.47	3.48
		iː	531	2225	3069	3.03	3.04
		iː	584	2086	3076	3.88	3.89
		iː	734	2225	3069	3.03	3.04
	distress	ɪ	501	2167	2266	10.04	10.04
		ɪ	668	2130	3126	8.36	8.37
	distress	ɛ	560	2061	3082	2.28	2.3
		ɛ	769	2275	2955	2.38	2.39
		ɛ	853	2082	3130	9.53	9.54
	distress	a	525	1059	1596	12.4	12.4
		a	566	1099	1644	13.43	13.44
		a	600	1118	1766	14.87	14.88
		a	653	1119	1785	17.41	17.41
		a	684	1513	2417	9.72	9.72
	distress	ɒ	702	1361	2508	11.36	11.36
		ɒ	715	1414	1962	16.76	16.77
		ɒ	781	1552	2491	7.31	7.32
		ɒ	817	1524	3235	1.38	1.39
	distress	ʌ	587	1133	1690	8.82	8.83
		ʌ	831	1292	2510	8.58	8.59
E	distress	iː	365	2108	2611	4.05	4.06
		iː	456	2522	2647	1.54	1.55
		iː	473	2465	2634	5.12	5.13
	distress	ɛ	728	1810	2998	0.15	0.16
	distress	ɒ	789	1259	2364	7.14	7.16
		ɒ	818	1256	2010	6.46	6.47
		ɒ	860	1238	2136	8.39	8.39
	distress	ʌ	712	1747	2441	7.88	7.89
		ʌ	770	1550	2500	3.78	3.79
F	distress	iː	427	1275	2312	4.11	4.12
		iː	510	1519	2607	6.02	6.03

F5: Mean formant values (Hz) for male actors across all conditions.

	F1			F2			F3		
Speaker	**ref**	**unreh**	**reh**	**ref**	**unreh**	**reh**	**ref**	**unreh**	**reh**
Act 1	**349**	**401**	**374**	**2230**	**2367**	**2203**	**2892**	**3112**	**2539**
iː	349	401	374	2230	2367	2203	2892	3112	2539
Act 2	**319**	**345**		**1996**	**2147**		**2392**	**2646**	
iː	319	345		1996	2147		2392	2646	
Act 3	**454**	**514**	**507**	**1643**	**1580**	**1535**	**2516**	**2441**	**2530**
iː	332	355	326	2157	2132	2105	2690	2696	2431
ɪ	414	389	434	1653	1667	1722	2415	2488	2714
a	560	676	705	1494	1379	1288	2511	2409	2497
ʌ	510	638	561	1268	1141	1026	2448	2172	2480
Act 4	**534**	**575**	**532**	**1640**	**1726**	**1636**	**2703**	**2585**	**2590**
iː	304	382	394	2142	2179	2277	2863	2932	3668
ɪ	436	484	495	1704	1920	1902	2489	2349	2401
a	741	783	746	1550	1650	1513	2761	2613	2397
ʌ	657	652	493	1163	1154	853	2699	2445	1895
Act 5	**533**	**587**	**574**	**1635**	**1758**	**1771**	**2706**	**2645**	**2681**
iː	328	361	320	2115	2179	2173	2936	2775	2757
ɪ	431	473	470	1745	2053	2032	2581	2570	2758
ɛ	582	685	625	1670	1839	1782	2713	2709	2642
a	712	758	748	1460	1506	1536	2554	2727	2417
ʌ	612	658	706	1187	1215	1332	2745	2447	2832
Act 6	**502**	**510**	**502**	**1693**	**1776**	**1635**	**2610**	**2640**	**2460**
iː	293	327	335	2277	2270	1999	2823	2891	2534
ɪ	377	363	397	1772	1965	1867	2584	2583	2573
ɛ	561	549	546	1672	1784	1597	2648	2713	2427
a	664	725	692	1431	1559	1413	2427	2498	2289
ʌ	612	587	540	1312	1303	1302	2568	2516	2477

F6: Mean formant values (Hz) for female actors across all conditions.

	F1			F2			F3		
Speaker	**ref**	**unreh**	**reh**	**ref**	**unreh**	**reh**	**ref**	**unreh**	**reh**
Act 7	**667**		**696**	**1733**		**1739**	**2859**		**2802**
iː	380		426	2686		2114	3223		2665
ɪ	519		500	1993		1885	2826		2975
ɛ	649		697	1893		1955	2860		2826
a	860		938	1639		1861	2668		3333
ɑː	784		741	1261		1301	2686		2324
ɒ	668		802	1172		1329	2810		2705
ʌ	810		765	1487		1726	2945		2789
Act 8	**634**	**713**	**697**	**1767**	**1793**	**1833**	**2906**	**2562**	**2931**
iː	350	431	397	2617	2861	2698	3202	3446	3415
ɪ	470	553	565	2056	1873	2090	3053	2957	3116
ɛ	649	891	788	1886	1562	1920	2898	2049	2865
a	951	849	945	1725	1684	1716	3070	2124	2684
ɑː	661	788	694	1259	1623	1541	2370	2652	3213
ɒ	653	698	760	1393	1288	1429	2856	2321	3027
ʌ	701	784	731	1436	1662	1437	2893	2388	2200
Act 9	**609**	**648**	**626**	**1830**	**1723**	**1710**	**2997**	**2886**	**2941**
iː	408	428	485	2647	2402	2456	3253	3019	3305
ɛ	663	732	817	2002	1947	1983	3071	3061	2811
ɒ	620	672	569	1202	1181	1021	2821	2786	2927
ʌ	743	760	631	1470	1362	1382	2843	2676	2722
Act 10	**588**	**558**	**538**	**1698**	**1678**	**1555**	**2961**	**2393**	**2487**
iː	343	360	325	2552	2325	2198	3406	2684	3136
ɛ	648	668	622	1843	1967	1666	2945	2665	2852
ɒ	698	609	690	1128	1210	1306	2695	2131	2322
ʌ	665	596	515	1270	1211	1051	2799	2093	1640
Act 11	**400**	**420**	**353**	**2619**	**2840**	**2730**	**3114**	**3311**	**3479**
iː	400	420	353	2619	2840	2730	3114	3311	3479
Act 12	**342**	**369**	**486**	**2328**	**2222**	**2294**	**2860**	**2533**	**3109**
iː	342	369	486	2328	2222	2294	2860	2533	3109

Appendix G - Perceptual Experiment (Chapter 6)

G1 Information Sheet

Perceptual Experiment on Distress Speech

What is the experiment about?
This experiment looks at acted and real-life distress to see if some audio extracts of distress sound more genuine than others.

Why have I been invited to take part?
You have been asked to take part in this experiment as you have expressed a willingness to be involved in research about genuine sounding distress. The experiment will help find out what the average person thinks genuine distress sounds like.

What is involved?
You will be asked to listen to extracts from acted and real 999 calls. These calls contain references to violent crime and may include descriptions of violence. Some involve people who have been physically and violently attacked and others involve actors who are pretending to be attacked. You will hear extracts of varying durations from a variety of voices – e.g. male/female, old/young, Northerner/Southerner etc. Some extracts may contain the same words but that doesn't mean they're from the same person you may have heard earlier - 999 calls are often formulaic so some of the same words occur in different calls. Each extract will be played twice with a pause in between. You will hear a double beep at the start of each new extract and a single beep before the second repetition. You will then be asked to determine whether the extract was produced by an actor or a victim. You will also be asked how confident you are in making that assessment. There is space on the response sheet for you to write down anything noticeable about the extract that you think may be important, if you want to. Please note that the extracts recorded by actors are designed to mimic 999 calls so they may sound like they are of bad quality (which is typical of these sorts of calls). The experiment will take about 20 minutes. All response sheets remain anonymous.

Do I have to take part?
No. If you decide not to take part, that's OK. If you do decide to take part, you will be asked to sign a consent form before the experiment. If you start the experiment and then later change your mind, you can withdraw at any time. You do not need to give a reason. If you decide to stop the experiment, your response sheet will be destroyed.

How will the information I provide be used?
The responses will be looked at by Lisa to see which extracts are correctly identified as being produced by victims and which are correctly identified as being produced by actors. Where extracts are found to belong to victims, she will analyse the call it came from to see what sort of features may be contributing to making it sound more genuine. This information will be useful for speech researchers and drama teachers to see which aspects of acted and real distress performances are perceived as portraying more distress than others by audiences.

Based on the findings, Lisa may present the research in scholarly reports and conference presentations. She is happy to keep you informed of these research findings. Please indicate on the consent form if you would like Lisa to email you the findings from the experiment once the research has been completed.

What are the possible benefits and risks of taking part?
It is hoped that you will find the experience rewarding and stimulating. Please bear in mind that, given the nature of 999 calls, some people may find the extracts upsetting. If you think you are likely to find this distressing, or if you are unsure, please do not take part in the experiment. If at any time you would like to stop taking part in the experiment, that's OK. You are under no obligation to continue taking part in the experiment if you don't want to. You can stop at any time, no questions asked. If you do stop during the experiment, your response sheet will be destroyed.

Who is the running the research?
The experiment is being run by Lisa Roberts, a PhD student at the University of York, in the Department of Language and Linguistic Science. Her research is being supervised by Prof. Peter French and Prof. Paul Foulkes.

Lisa's contact details
Department of Language and Linguistic Science
University of York
Heslington
York Y010 5DD
Telephone: 01904 432650
Email: lsr501@york.ac.uk

G2 Consent Sheet

Consent to Participate in Research

I agree to take part in this experiment, for which I have volunteered because I am comfortable to help research which investigates our perception of distress, such as that heard in 999 calls, which may include descriptions of violence and involve victims of violent crime.

I acknowledge that I understand:

- what the experiment involves
- what the research is about
- that my name and any details will be protected
- that I'm willing to hear both genuine and acted 999 calls of a potentially distressing situation

I understand that I can withdraw my participation at any time, and that I am under no obligation to complete the experiment. If I do stop the experiment, my response sheet will be destroyed.

Any questions I had about the study have been answered.

I would/would not like to hear about the results of this experiment via email once the research has been completed.
Please indicate your email address here if yes ……

Signed:_____

Name:_____

Date:_____

(You will be given a copy of this consent form for your records.)

G3 Response Sheet

Experiment Info

(please circle)

Experiment: A B

Participant Info

(please circle or state where applicable)

Sex: M F

Age: ……

Native Accent/place you grew up ……..
……..
……..

Level of education completed ……..
……..
……..

Have you ever had to call 999? Y N

If Yes, please give details: ……..
……..
……..
……..
……..
……..

Have you ever heard a 999 call before? Y N

If Yes, please give details: ……..
……..
……..
……..
……..
……..

Extract 01:

1. Do you think the speaker in this extract is a real victim or an actor? *(Please circle)*

Definitely victim	*Probably victim*	*No decision*	*Probably actor*	*Definitely actor*

2. How certain are you? *(Please circle)*

Very certain	*Quite certain*	*Neither certain nor uncertain*	*Quite uncertain*	*Very uncertain*

3. Do you think the person is male or female? *(Please circle)*

Male	*Female*	*No decision*

4. Notes:
(Please feel free to make notes if there's anything about this extract you think might be important)

Extract 02:

1. Do you think the speaker in this extract is a real victim or an actor? *(Please circle)*

Definitely victim	*Probably victim*	*No decision*	*Probably actor*	*Definitely actor*

2. How certain are you? *(Please circle)*

Very certain	*Quite certain*	*Neither certain nor uncertain*	*Quite uncertain*	*Very uncertain*

3. Do you think the person is male or female? *(Please circle)*

Male	*Female*	*No decision*

4. Notes:
(Please feel free to make notes if there's anything about this extract you think might be important)

Extract 03:

1. Do you think the speaker in this extract is a real victim or an actor? *(Please circle)*

Definitely victim	*Probably victim*	*No decision*	*Probably actor*	*Definitely actor*

2. How certain are you? *(Please circle)*

Very certain	*Quite certain*	*Neither certain nor uncertain*	*Quite uncertain*	*Very uncertain*

3. Do you think the person is male or female? *(Please circle)*

Male	*Female*	*No decision*

4. Notes:
(Please feel free to make notes if there's anything about this extract you think might be important)

G4 List of extracts in perceptual experiment

Case	Speaker	Speech/ Scream	Extract	Pool	Content	Duration of extract (s)	Duration of speech material (s)
A	Vic A	speech	A	A	Oh. Please. No	4.7	1.4
A	Actor 1	speech	B	B	Oh. No. Don't you- Please. Don't	6.1	2
A	Actor 2	speech	C	B	Oh. No. Don't you- Please. Don't	10	3.9
A	Vic A	scream	D	B	[scream]	4.2	0.8
A	Actor 1	scream	E	A	[scream]	3.6	0.8
A	Actor 2	scream	F	A	[scream]	3.5	0.7
B	Vic B	speech	G	A	I've been stabbed	3.7	0.8
B	Actor 3	speech	H	B	I've been stabbed	4.1	1.1
B	Vic B	speech	I	B	Yeah. Like, I'm on- on the road, on the main road.	7.4	2.7
B	Actor 4	speech	J	A	I'm- I'm on the, I'm down on the, um I'm down on the	6.5	2.3
B	Vic B	speech	K	A	Co-, I think that I'm going to pass out, very quickly. Honestly, I'm bleeding like mad here.	8.9	3.4
B	Actor 4	speech	L	B	Come on, I think that I'm going to pass out, very quickly. Honestly, I'm bleeding like mad here.	13.2	5.6
C	Vic C	scream	M	B	[screaming]	8.3	3.2
C	Actor 5	scream	N	A	[screaming/vocalising/sobbing]	10.3	4.1
C	Actor 6	scream	O	A	[screaming]	9.3	3.7
C	Vic C	speech	P	A	I'm bleeding to death. I've been stabbed through the thro-	6.1	2.1
C	Actor 5	speech	Q	B	can't, I'm bleeding to death. I've been stabbed through the-	7.4	2.7
C	Vic C	speech	R	B	Yeah. I'm dead	3.6	0.9
C	Actor 6	speech	S	A	Yeah. I'm dead	2.9	0.5
D	Vic D	scream	T	A	No!	1.7	5.3
D	Actor 8	scream	Y	B	No!	1	3.9
D	Vic D	scream	W	B	[screaming]	3.9	1

D	Actor 7	scream	X	A	[screaming]	7.9	2.6
D	Vic D	speech	Z	A	Stop it. Get out. Leave me alone.	9.1	3.6
D	Actor 7	speech	AA	B	Stop it. What are you doing? Get out. Leave me alone.	8	2.9
D	Vic D	speech	AB	B	I've locked myself in the bathroom.	5.3	1.7
D	Actor 8	speech	AC	A	I've locked myself in the bathroom.	4.9	1.5
E	Vic E	speech	AD	B	I'm scared.	3.4	0.7
E	Actor 10	speech	AE	A	I'm scared.	5.4	1.6
E	Vic E	speech	AF	A	I've been shot.	4	1
E	Actor 9	speech	AG	B	I've been shot.	1.2	4.4
E	Vic E	speech	AH	B	Come on.	3.3	0.7
E	Actor 10	speech	AI	A	Come on.	3.7	0.9
F	Vic F	speech	AJ	B	[vocalisation]	3.5	0.8
F	Actor 11	speech	AK	A	Police. Please.	6.1	2.1
F	Actor 12	speech	AL	A	Please. Police	10.7	4.2
F	Vic F	scream	AM	A	[screaming]	4.1	1.1
F	Actor 11	scream	AN	B	[screaming/sobbing]	17.8	7.2
F	Actor 12	scream	AO	B	[screaming]	16.6	7.3
F	Actor 12		AP	control	[vocalisation] No [vocalisation]	14.9	5.7
C	Vic C		AQ	control	Yeah [vocalisation]	7	2.6
B	Actor 3		AU	control	Come on. I'm bleeding like mad here.	6.7	2.3
D	Vic D		AS	control	He's still got a gun in his-	1.7	5.4

G5 Table of comments volunteered by participants taking part in the perceptual experiment

Extract	Group	Comment
A	LAY	too clear and the scream sounded misplaced somehow
	LAY	More difficult for men to sound so emotional (genuinely)
AA	FORENSIC	phrases follow too fast from one another, as if reading from a script
	FORENSIC	Sounded like a case I've encountered though the recording sounded clearer than I expect from on outdoor recording which this seemed to be
	FORENSIC	High intelligibility/articulateness a significant cue
	LAY	no space for answering comments - too measured. Sounds like a quarrel, not an emergency
	LAY	Just sounds "unrealistic" but I can't explain why.
	LAY	Quite eloquent!
	POLICE	Words too well pronounced
AB	FORENSIC	sounds urgent and rushed
	FORENSIC	Gender judgments seem to be easier. I haven't found a clear rationale for indentifying an actor yet
	FORENSIC	breathiness and tempo fairly convincing but other factors could be inconsistent
	LAY	too calm
	LAY	Northern
AC	FORENSIC	implausible content for fabricated 999
	LAY	too calm? Not speaking quietly for someone hiding in a bathroom
AD	FORENSIC	very under-acted for an actor
	FORENSIC	I'm beginning to feel like I cannot find any useful markers
	FORENSIC	too brief for reliable judgement
	FORENSIC	too short
	LAY	More calmness, so I think is genuine.
	POLICE	Sounds automated
	POLICE	English not first language
AE	FORENSIC	didn't sound all that scared, contrary to the words
	FORENSIC	judgement largely based upon voice quality
	LAY	do people actually say "I'm scared" when they're really afraid?
AF	FORENSIC	not dramatic enough to be an actor
	FORENSIC	sounds neutral, not stressful
	LAY	can't hear the words - sounds calm but urgent
	LAY	Weirdly calm
	LAY	seems too unconcerned
	LAY	Too calm
	POLICE	female
	POLICE	Shot

	POLICE	Too calm
AG	FORENSIC	loud in breath again. Not exactly critical I guess but I am grasping at straws
	FORENSIC	Plausibly authentic; brevity reducing certainty
	LAY	far too calm
	LAY	Sounds like a line from the film "Lock, Stock, and Two Smoking Barrels". i.e. "Would everyone stop getting shot!"
	POLICE	Doesn't sound distressed enough to of [sic] been shot
	POLICE	Sounded in need of heLay
AH	FORENSIC	too short
	FORENSIC	too brief
	FORENSIC	too short
	LAY	nothing to go on
AI	FORENSIC	too short to tell
	FORENSIC	too brief
	FORENSIC	too short
	POLICE	distressed
AJ	FORENSIC	too short
	FORENSIC	What kind of noise is that to make if distressed?
	FORENSIC	! Sounds like a bird cry
	FORENSIC	too brief
	FORENSIC	very short, taped over her mouth?
	LAY	Not enough to go on.
AK	FORENSIC	had a nice dramatic regularity to it
	FORENSIC	rattle sound
	FORENSIC	breathing voice quality/tempo + recording quality acted as cues
	FORENSIC	child?
	LAY	didn’t sound pleading despite the words spoken
	LAY	Voice - full of emotion
	POLICE	Had convincing quiver in her voice
	POLICE	Sounded like begging for help
AL	FORENSIC	very brief, too short to tell
	FORENSIC	too little speech to be certain
	FORENSIC	breathing pattern + pitch significant cues
	LAY	Difficulty in fighting to control emotions seems genuine
	POLICE	cannot tell from the shortness
AM	FORENSIC	very short; no intelligible speech
	FORENSIC	certainty non-judgement owing to brevity
	LAY	too weird a sound to fake
	POLICE	Very weird sound
	POLICE	Sounds like she's on a rollercoaster
AN	FORENSIC	gasps outside the speech either very good acting, or genuine
	FORENSIC	desperate (by a good actor?)

	FORENSIC	breath patterning a significant cue
	FORENSIC	if not, extremely good actress. Sounds convincing
	LAY	Again to me it's the breathing that makes it sound real
	LAY	She's saying "No" in disbelief, it sounds like as if she can't believe she's been stabbed/raped/…
AO	FORENSIC	was this all on inspired breath? I have really no idea if this was real or acted but sounded deliberate so I switched decision.
	FORENSIC	(I'm starting to think there's no way of telling, given a good actor)
	FORENSIC	Voice quality and pitch fairly authentic
	FORENSIC	highly unusual for an actor, in movies etc
	FORENSIC	too much
	LAY	Old woman? I can't believe these sounds would be made by anyone
AP	FORENSIC	"no" sounded stress-free and broke the mood
	FORENSIC	"no" to normal for someone in distress
	FORENSIC	Timing sounded too deliberate - but I may well be wrong
	FORENSIC	the noise he makes doesn't sound like someone in real pain
	FORENSIC	to parts to stimulus, each producing separate judgements
	FORENSIC	if victim, probaby very drunk (or on drug), but room acoustics influence me to think actor…
	FORENSIC	I see we get them again. Can't recall what I thought last time. Oh yes, rather deliberate.
	FORENSIC	internally contradictory (voice quality)
	LAY	sounds like someone after a heavy lunch
	LAY	Seems 'studied', rather artificial
	LAY	just sounded fake -the groans didn't sound in sync with the "no"
	LAY	The 3rd sound "Ahhh!" sounded really fake
	LAY	Doesn't sound v natural!
	LAY	Another repeat
	POLICE	Seemed put on groaning
	POLICE	Not sure
	POLICE	Likely drunk
	POLICE	Sounded a bit forced
	POLICE	Appeared to be 2 very different sounds
	POLICE	Elderly
	POLICE	Sounds like extract 2
AQ	FORENSIC	quite regular steps in the ?falsetto
	FORENSIC	quite regular steps in the falsetto
	FORENSIC	ah, are they repeating?
	FORENSIC	I can't recall encounting sounds like this on my case work so it feels like I am making a rather random decision
	FORENSIC	sounds involuntary, weird, feminine; not typical of actors
	FORENSIC	short
	FORENSIC	Haven't we heard this before?

	FORENSIC	impossible to discriminate between histrionic performance and what might be considered the "real thing"
	FORENSIC	can't remember what I judged the last time…
	LAY	Rehearsed
	LAY	the break in the voice sounded realistic
	LAY	This is a repeat of an earlier extract
	POLICE	Sounds like elderly male
AS	FORENSIC	very strained phonation
	FORENSIC	sounds like one I heard in the beginning
	FORENSIC	something made me doubt it was a real call but really I am so uncertain. I marked the loud inbreath at the start
	FORENSIC	somewhat brief for any reliable decision
	FORENSIC	too short
	FORENSIC	too short
	FORENSIC	Loud in breath again so shouldn't this be actor? Possibly but more likely it means I still can't work out a decision process
	FORENSIC	(Have I heard this one already?)
	FORENSIC	too brief; would have listened again if possible
	LAY	could not understand the words
	LAY	Genuine because began instantly hysterical
	LAY	sounded genuinely stressed
	LAY	I couldn't make out any words
	POLICE	Sounded distressed (female)
	POLICE	Sounded distressed but couldn't understand the words being shouted
	POLICE	Shouted, sounded desperate
AU	FORENSIC	I can't recall encountering sounds like this on my case work so it feels like I am making a rather random decision
	FORENSIC	brevity/quality make a judgement problematic
	FORENSIC	the crying in her voice sounds convincing to me
	FORENSIC	(Have I heard this one already?)
	FORENSIC	earlier sample
	LAY	sounds like the caller is trying to be calm though their fear
	LAY	couldn't hear the words
	LAY	Quite high pitched for a male voice. I would think that would be quite realistic, but this wasn't I felt.
	LAY	Can't tell…
	POLICE	Sounded like elderly person
	POLICE	female
	POLICE	Sounded genuinely upset
B	FORENSIC	unable to derive much from this sample
	FORENSIC	couldn't make out what was being said
	POLICE	Young 20/25
C	FORENSIC	echo, as if on stage; too regular, and not much feeling
	FORENSIC	sounded stilted

	FORENSIC	recording quality (reverberant) suggesting authenticity; judgement not based on voice (exclusively)
	LAY	Too formulaic to be realistic. "No". "Please". "Don't" Sounded scripted
	LAY	Pauses seem artificial
D	FORENSIC	too short to know; definitely sounds distressed
	FORENSIC	if anything, female. Sounds like a circular saw
	FORENSIC	Voice quality the main cue
	LAY	Was this done in a recording room?
E	FORENSIC	too short to tell
	FORENSIC	sounded quite controlled
	FORENSIC	sounds like a previous speaker (I think, but not sure)
	FORENSIC	too brief
	LAY	just not enough to go on. Couldn't say anything with confidence
	POLICE	Not sure why
	POLICE	Too short to decide
F	FORENSIC	too short to tell
	FORENSIC	an echoey theatre? Sample too short to know really
	FORENSIC	too short for decision
	LAY	Not a lot to go on!
	POLICE	Too short
G	FORENSIC	too matter-of-fact for an actor
	FORENSIC	no distress
	LAY	far too calm - no pain in the voice for someone who'd been stabbed
	LAY	Strangely controlled - could be either
	POLICE	not real
	POLICE	Male was a bit "matter-of-fact"
H	FORENSIC	(I wonder whether the echo made em think it was staged?)
	LAY	no pain in his voice
	LAY	Doesn't sound like he's been stabbed
I	FORENSIC	not acted - hesitation/repetition too natural
	FORENSIC	Non fluencies sounded 'natural' but nothing an actor couldn't do I imagine. There was no traffic noise so maybe it was an actor after all!
	FORENSIC	Articulation rate a salient cue
	LAY	sounded like genuine confusion and lack of clarity
	LAY	Bizarrely, calmness suggests this is genuinely
	POLICE	Sounded a bit shaky
J	FORENSIC	hesitation pattern unlikely to be replicated by an actor
	FORENSIC	I believe I thought the hesitation somehow convincing. I'm less convinced that I'm convinced.
	FORENSIC	dysfluency quite compelling
	FORENSIC	couldn't make out what he was saying
	LAY	stuttering and lack of clarity as to location seemed genuine

	LAY	Hesitation seems genuine
	LAY	Unsteady, confused speech
	POLICE	Sounded rehearsed
	POLICE	Does not sound real
	POLICE	Sounded nervous (but sounded forced)
K	FORENSIC	has the matter-of-fact tone that an actor wouldn't think of using
	FORENSIC	It sounds so matter-of-fact, it's hard to imagine an actor playing it like that
	LAY	sounded like someone a bit over-dramatic but somehow genuine
	LAY	Trying to give coherent info = genuine
	LAY	It's got that "Crimewatch" sound quality feel.
L	FORENSIC	breathing realistic
	FORENSIC	Sounded natural particularly at the end with apparent laugh and speech
	FORENSIC	his voicing may not be in line with his claimed health situation
	LAY	he words didn't fit with the voice - the last version of these words sounded more real
M	FORENSIC	nice an regular - + I think no sign of distressed breathing between the phonation
	FORENSIC	Just sounded desperate enough to be real - or a good actor!
	FORENSIC	voice quality the salient cue
	FORENSIC	not typical for movie actor, maybe a real stage actor? :) Could also be genuine
N	FORENSIC	sounds like acting
	FORENSIC	echo, reverb sounds unnatural
	FORENSIC	breath pattern difficult to reconcile with acted voice
	LAY	Regrettably I think this is another rape victim
O	FORENSIC	gasping for breath unlikely to be replicated by actor
	FORENSIC	The ingressive one. Is this really about acting or male vs female.
	FORENSIC	it's the hiccup that does it
	LAY	The whimper at the end didn't sound acted
	POLICE	A bit over dramatic
P	POLICE	Does not sound real
Q	FORENSIC	I thought victim but is it just because there is more speech. I really don't know.
	FORENSIC	utterance final breathing pattern is main determinant of judgement
	LAY	can't hear the words
	POLICE	Sounded scared and like she was crying
R	FORENSIC	too short
	FORENSIC	unable to extract much relevant information
	FORENSIC	too short
S	FORENSIC	too short to tell
	FORENSIC	underplayed for an actor

	FORENSIC	too short
	FORENSIC	too brief
	FORENSIC	too short
	LAY	I think they said "I'm dead" - but didn't sound scared or in pain
	LAY	Tense 'tone' in voice
	POLICE	Too short to hear anything
T	FORENSIC	too short
	FORENSIC	sounded genuinely distressed; maybe just good acting
	FORENSIC	shouting quality + pitch profile swayed decision
	LAY	that scream was strange and would be a strange sound to act/fake
	POLICE	Voice, sound
W	FORENSIC	too short to tell
	FORENSIC	too short + poor quality really to judge
	FORENSIC	too short
	FORENSIC	Another ingressive air stream
	FORENSIC	Too short to judge - not 100% sure it's a voice
	FORENSIC	pitch and voice quality suggest authenticity
	FORENSIC	too short
	LAY	Just not enough to go on.
X	FORENSIC	sounds a bit formulaic
	LAY	strange scream. Just didn't sound real
	LAY	No 'depth' to scream
Y	FORENSIC	If acted, a very competent thespian. Resonance properties of recording adds to 'authenticity'
	FORENSIC	loud, but no real distress
	LAY	Voice sounded realisitic but I'm not convinced by just shouting the word "No"
	LAY	Child?
	POLICE	a bit extreme to be genuine
Z	FORENSIC	shouted voice quality perceptually plausible
	LAY	too controlled and well-enunciated to be real
	LAY	Too 'coherent' to be genuine
	LAY	The echo of this room makes it sound like a real place.
	POLICE	Sounded authentic
	POLICE	Late teens/early 20's

References

Akwagyira, A. (2005). Does 'happy slapping' exist? *BBC News*. [online] 12 May 2005. Available at: http://news.bbc.co.uk/1/hi/uk/4539913.stm. [Accessed 01 October 2011]

Bachorowski, J.A. & M. J. Owren (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *Journal of the Acoustical Society of America* 106: 1054-1063.

Bachorowski, J.-A. & M. J. Owren (2001). Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science* 12: 252-257.

Baken, R. J. (1987). *Clinical Measurement of Speech and Voice*. London: Taylor & Francis Ltd.

Banse, R. & K. R. Scherer (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* 70: 614-636.

Barkhuysen, P., E. Krahmer & M. Swerts (2007). Cross-modal perception of emotional speech. *Proceedings of the 16th International Congress of Phonetic Sciences*.Saarbrücken, Germany. 2133-2136.

Begault, D. R. (2008). Forensic analysis of the audibility of female screams. *Proceedings of the Audio Engineering Society 33rd International Conference: Audio Forensics-Theory and Practice*. Denver, Colorado. 67-71.

Belin, P., S. Fillion-Bilodeau & F. Gosselin (2008). The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods* 40: 531-539.

Belin, P., R. J. Zatorre & P. Ahad (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research* 13: 17-26.

Benedetti, J. (2000). *Stanislavski: An Introduction, Revised and Updated*. New York: Routledge.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences (IFA)* 17, Amsterdam, Netherlands. 97-110

Boersma, P. & D. Weenink (2009). Praat: doing phonetics by computer. [Computer program]. Version 5.1.25.Available at: http://www.praat.org/ [Accessed 01 October 2009].

Bonebright, T. L., J. L. Thompson, & D. W. Leger (1996). Gender stereotypes in the expression and perception of vocal affect. *Sex Roles* 34: 429-445.

Brenner, M., T. Shipp, E. T. Doherty & P. Morrissey (1983). Voice measures of psychological stress - laboratory and field data. *In* I. R. Titze & R. C. Scherer (eds.) *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*. Denver, Denver Center for Performing Arts. 239-248.

Brown, C., F. Alipour, D. A. Berry. & D. Montequin (2003). Laryngeal biomechanics and vocal communication in the squirrel monkey (Saimiri boliviensis). *The Journal of the Acoustical Society of America* 113 2114-2126.

Burkhardt, F. & W. F. Sendlmeier (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. *Interational Speech Communication Association (ISCA) Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK.

Byrne, C. & P. Foulkes (2004). The 'Mobile Phone Effect' on vowel formants. *International Journal of Speech, Language and the Law* 11: 83-102.

Caffi, C. & R. W. Janney (1994). Toward a pragmatics of emotive communication. *Journal of Pragmatics* 22: 325-373.

Cambier-Langeveld, T. (2010). The role of linguists and native speakers in language analysis for the determination of speaker origin. *International Journal of Speech, Language and the Law* 17: 67-93.

Chin, S. B. &D. B. Pisoni (1997). *Alcohol and Speech*. San Diego: Academic Press.

Chung, S. (2000). *Expression and Perception of Emotion extracted from the Spontaneous Speech in Korean and English*. Unpublished: Sorbonne Nouvelle University. PhD.

Clifford, B. R., H. Rathborn & R. Bull (1980). The effects of delay on voice recognition accuracy. *Law and Human Behavior* 5: 201-208.

Cornelius, R. R. (1996). *The Science of Emotion. Research and Tradition in the Psychology of Emotion*. Upper Saddle River, New Jersey: Prentice-Hall.

Coupland, N. & H. Bishop (2007). Ideologised values for British accents. *Journal of Sociolinguistics* 11(1): 74-93.

Cowie, R. & R. R. Cornelius (2003). Describing the emotional states that are expressed in speech. *Speech Communication* 40: 5-32.

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz & J.G. Taylor (2001). Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE* 18(1): 32-80.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Darwin, C. R. (1872). *The Expression of The Emotions in Man and Animals*. London: John Murray.

David, M. (1995). *The New Voice Pedagogy*. Lanham, Maryland: Scarecrow Press.

Davidson, R. J., K. R. Scherer & H. Goldsmith (2003). *Handbook of Affective Sciences*. Oxford: Oxford University Press.

Demolin, D. (2007). Phonological universals and the control and regulation of speech production. *In* P. Beddor, M. Ohala and M. J. Solé (eds.) *Experimental Approaches to Phonology*. Oxford, Oxford University Press. 75-92.

Devillers, L., I. Vasilescu & L. Vidrascu (2004). F0 and pause features analysis for anger and fear detection in real-life spoken dialogs. *Proceedings of Speech Prosody 2004*, Nara, Japan. 205-208.

Doscher, B. M. (1994) *The Functional Unity of the Singing Voice*. Lanham, Maryland: Scarecrow Press.

Douglas-Cowie, E., Campbell, N., Cowie, R. & Roach, P. (2003). Emotional Speech: Towards a new generation of databases. *Speech Communication* 40: 33-60.

Douglas-Cowie, E., Cowie, R. & Schröder, M. (2000). A new emotion database: Considerations, sources and scope. *International Speech Communication Association (ISCA) Tutorial and Research Workshop (ITRW) on Speech and Emotion.* Newcastle, Northern Ireland, UK. 39-44.

Edwards, J. & M. Jacobsen (1987). Standard and regional standard speech: distinctions and similarities. *Language in Society* 16(3): 369-379.

Ekman, P. (1971). Universals And Cultural Differences In Facial Expressions Of Emotions. *Proceedings of the Nebraska Symposium on Motivation*, Lincoln, Nebraska: University of Nebraska Press. 207-283

Ekman, P. (1992). An Argument For Basic Emotions. *Cognition and Emotion* 6: 169-200.

Ekman, P. (1994). All emotions are basic. *In* P. E. Ekman and R. J. Davidson (eds.) *The nature of emotion: Fundamental questions*. New York: Oxford University Press. 15-19.

Ekman, P., W. V. Friesen & K. R. Scherer (1976). Body movement and violence pitch in deceptive interaction. *Semiotica* 16: 23-27.

Elliot, N., J. Sundberg & P. Gramming (1995). What happens during vocal warm-up? *Journal of Voice* 9: 37-44.

Eriksson, A. (2005). Tutorial on Forensic Speech Science. *Interspeech (9th European Conference on Speech Communication and Technology).* Lisbon, Portugal. Available at: https://www.york.ac.uk/media/languageandlinguistics/documents/currentstudents/Eriksson_tutorial_paper.pdf [Accessed 30 September 2012]

Fant, G. (1971). *Acoustic Theory of Speech Production: With Calculations Based on X-ray Studies of Russian Articulations*. Walter de Gruyter.

Fecher, N. & D. Watt (1989). Speaking under cover: The effect of face-concealing garments on spectral properties of fricatives. *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong,China. 663-666.

Fischer, J., K. Hammerschmidt, D. L. Cheney & R. M. Seyfarth (2002). Acoustic features of male baboon loud calls: influences of context, age, and individuality. *The Journal of the Acoustical Society of America* 111: 1465-1474.

Foulkes, P. & G. J. Docherty (2006). The social life of phonetics and phonology. *Journal of Phonetics* 34: 409-438.

Fraser, H. (2003). Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases. *International Journal of Speech, Language and the Law* 10: 203-226.

Fraser, H. (2009). The role of 'educated native speakers' in providing language analysis for the determination of the origin of asylum seekers. *International Journal of Speech, Language and the Law* 16: 113-138.

.

French, J. P. (1990). Analytic procedures for the determination of disputed utterances. *In* H. Kniffka ed. *Texte zu Theorie und Praxis forensischer Linguistik.* Tübingen, Max Niemayer Verlag. 201-213.

Fry, D. B. (1979). *The Physics of Speech*: Cambridge: Cambridge University Press.

Fuller, B. F., Y. Horii & D. A. Conner (1992). Validity and reliability of nonverbal voice measures as indicators of stressor-provoked anxiety. *Research in Nursing & Health* 15: 379-389.

Garnier, M., N. Henrich, J. Smith & J. Wolfe (2010a). The tuning of vocal resonances and the upper limit to the high soprano range. *International Symposium on Music Acoustics (ISMA 2010)*, Sydney and Katoomba, Australia. 11-16

Garnier, M., N. Henrich, J. Smith & J. Wolfe (2010b). Vocal tract adjustments in the high soprano range. *The Journal of the Acoustical Society of America* 127: 3771-3780.

Gebhard, P., M. Schröder, M. Charfuelan, C. Endres, M. Kipp, S. Pammi, M. Rumpler, & O. Türk (2008). IDEAS4Games: building expressive virtual characters for computer games. *Proceedings of Intelligent Virtual Agents (IVA)*, Tokyo, Japan. 426-440.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. New York: Academic Press.

Harrison, P. (2001). GSM interference cancellation for forensic audio: a report on work in progress. *International Journal of Speech, Language and the Law* 8: 9-23.

Harrison, P. (2004). *Variability of Formant Measurements.* Unpublished.University of York. MA Thesis.

Hess, U. & P. Thibault (2009). Darwin and emotion expression. *American Psychologist* 64: 120-128.

Hicks, J. W. Jr. (1979). *An Acoustical/Temporal Analysis of Emotional Stress in Speech.* Unpublised. University of Florida. PhD.

Hirson, A., J. P. French & D. Howard (1995). Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics. *In* J. Windsor-Lewis ed. *Studies in*

General and English Phonetics. Essays in Honour of J. D. O'Connor. London/New York, Routledge. 230-240.

Hirson, A. & D. M. Howard (1994). Spectrographic analysis of a cockpit voice recorder tape. *Forensic Linguistics* 1 59–69.

Hollien, H. (1980). Vocal indicators of psychological stress. *Annals of the New York Academy of Sciences* 347(1) 47-72.

Hollien, H., K. Liljegren, C. A. Martin & G. De Jong (2001). Production of intoxication states by actors - acoustic and temporal characteristics. *Journal of Forensic Science* 46: 68-73.

Honigsbaum, M. (2005). Concern over rise of 'happy slapping craze. *The Guardian.* [online] 26 April 2005. Available at: http://www.theguardian.com/uk/2005/apr/26/ukcrime.mobilephones. [Accessed 01 October 2011]

Huggins, A. W. (1980). Better spectrograms from children's speech: A research note. *Journal of Speech and Hearing Research* 23 19-27.

Husler, F. & Y. Rodd-Marling (1965). *Singing: The Physical Nature of the Voice Organ; a Guide to the Unlocking of the Singing Voice*. London: Faber and Faber.

IAFPA (2004). International Association of Forensic Phonetics and Acoustics (IAFPA) Code of Practice.[Online] Available at: http://www.iafpa.net/code.htm. [Accessed 01 September 2008]

Izard, C. E. (1971). *The Face of Emotion*. East Norwalk, Connecticut: Appleton-Century-Crofts.

Izard, C. E. (1977). *Human Emotions*. New York: Plenum Press.

Jessen, M. (1997). Phonetic manifestations of cognitive and physical stress in trained and untrained police officers. *Forensic Linguistics* 4: 124-147.

Jessen, M. (2006). *Einfluss von Stress auf Sprache und Stimme. Unter Besonderer Beruecksichtigung Polizeidienstlicher Anforderungen.* Idstein: SchulzKirchner Verlag GmbH.

Jessen, M. (2009). Forensic phonetics and the influence of speaking style on global measures of fundamental frequency. *In* G. Grewendorf and M. Rathert eds. *Formal Linguistics and Law*. New York: Walter de Gruyter. 115-140.

Jessen, M., O. Koster & S. Gfroerer (2007). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law* 12(2): 174-213.

Johnstone, T. (2001). *The effect of emotion on voice production and speech acoustics*. Unpublished. University of Western Australia. PhD

Johnstone, T. & K. R. Scherer (2000). Vocal communication of emotion. *In* M. Lewis and J. Haviland (eds). *Handbook of Emotion*. New York, Guilford. 220-235.

Jongman, A., Y. Wang, C. Moore & J. A. Sereno (2006). Perception and production of Mandarin Chinese tones. *In* E. Bates, L. H. Tan and O. Tseng (eds.) *Handbook of Chinese Psycholinguistics*. Cambridge, Cambridge University Press. 250-256.

Junqua, J. C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication* 20: 13-22.

Juslin, P. N. & K. R. Scherer (2005). Vocal expression of affect. *In* J. Harrigan, R. Rosenthal and K. Schere (eds.) *New Handbook of Methods in Nonverbal Behavior Research*. Oxford: Oxford University Press. 65-136.

Kennedy, G. A. (1972). *The Art of Rhetoric in the Roman world, 300 BC-AD 300*. Princeton, New Jersey: Princeton University Press.

Khattab, G. & J. Roberts (2010). Working with Children. *In* M. Di Paolo and M. Yaeger-Dror (eds.) *Sociophonetics: A Student's Guide*. Routledge. 163-178.

Kirchhübel, C. & D. M. Howard (2013). Detecting suspicious behaviour using speech: Acoustic correlates of deceptive speech - an exploratory investigation. *Applied Ergonomics* 44(5): 694-702.

Kirchhübel, C., D. M. Howard & A. Stedmon (2011). Acoustic correlates of speech when under stress: Research, methods and future directions. *International Journal of Speech, Language and Law* 18: 75-98.

Klasmeyer, G. & W. F. Sendlmeier (1997). The classification of different phonation types in emotional and neutral speech. *Forensic Linguistics* 4: 104-124.

Kuroda, I., O. Fuhiwara, N. Okamura & N, Utsuki (1976). Method for determining pilot stress through analysis of voice communication. *Aviation, Space and Environmental Medicine:* 47: 528-533

.

Künzel, H. J. (1997). Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech, Language, Law* 4: 48-83.

Künzel, H. J. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics* 8: 80-99.

Labov, W. (1994). *Principles of Linguistic Change, Vol. 1: Internal factors*. Oxford: Blackwell.

Lass, N. J., K. R. Hughes, M. D. Boyer, L. T. Waters & V.T. Bourne (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *Journal of the Acoustical Society of America* 59: 675-678.

Latinus, M. & P. Belin (2011). Human voice perception. *Current Biology* 21: 143-145.

Laukka, P., P. Juslin & R. Bresin (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion* 19(5): 633-653.

Laukka, P., C. Linnman, F. Åhs, A. Pissiota, Ö. Frans, V. Faria, Å. Michelgård, L. Appel, M. Fredrikson, & T. Furmark (2008). In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech. *Journal of Nonverbal Behavior* 32(4): 195-214.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Laver, J. (1994). *Principles of Phonetics*. Cambridge: Cambridge University Press.

Lazarus, R. S. (1966). *Psychological Stress and the Coping Process*. New York: McGraw-Hill.

Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist* 46: 819-834.

Lazarus, R. S., J. R. Averill & E. M. J. Opton (1970). Toward a cognitive theory of emotions. *In* M. Arnold (ed.) *Feelings and Emotions*. New York, Academic Press. 207-232.

Leinonen, L., T. Hiltunen, I. Linnankoski, & M.-L. Laakso (1997). Expression of emotional-motivational connotations with a one-word utterance. *Journal of the Acoustical Society of America* 102: 1853-1863.

McGlashan, J., C. Sadolin & H. Kjelin (2007). Can vocal effects such as distortion, growling, rattle and grunting be produced without traumatising the vocal folds? *Proceedings of the Pan European Voice Conference 7 (PEVOC 7)*. Groningen, The Netherlands.

McGlone, R. E. and W. H. Manning (1979). Role of the second formant in pitch perception of whispered and voiced vowels. *Folia Phoniatrica et Logopaedica* 31(1): 9-14.

Mehrabian, A. (1972). *Nonverbal communication*. Chicago: Aldine-Atherton.

Meinerz, C. (2008). Formant history, speech tempo and fundamental frequency under the influence of induced stress. Unpublished paper presented at the *International Association of Forensic Phonetics and Acoustics (IAFPA) Annual Conference 2008*. Lausanne, Switzerland.

Mende, W., H. Herzel and K. Wermke (1990). Bifurcations and chaos in newborn infant cries. *Physics Letters A* 145: 418-424.

Merlin, B. (2001). *Beyond Stanislavsky: The Psycho-physical Approach to Actor Training*. London: Nick Hearn Books.

Merlin, B. (2010). *Acting: The Basics*. New York: Routledge.

Merlin, B. (2014). *The Complete Stanislavsky Toolkit*. London: Nick Hern Books.

Milroy, L. & M. Gordon (2003). *Sociolinguistics: Method and Interpretation*. Oxford: Blackwell.

Moore, R. & I. Trancoso (1995). *Speech Under Stress: Proceedings of ESCA-NATO Tutorial and Research Workshop, Lisbon, Portugal 1995.* Lisbon, Portugal.

Morrison, D., R. Wang & L. C. De Silva (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication* 49(2): 98-112.

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist* 111: 855-869.

Motel, T., K. V. Fisher & C. Leydon (2003). Vocal warm-up increases phonation threshold pressure in soprano singers at high pitch. *Journal of Voice* 17(2): 160-167.

Murray, I. R. & J. L. Arnott (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication* 16(4): 369-390.

Murray, I. R., J. L. Arnott, N. Alm, & A.F. Newell (1991). A communication system for the disabled with emotional synthetic speech produced by rule. *Proceedings of the Second European Conference on Speech Communication and Technology (Eurospeech '91)*, Genova, Italy.

Murray, I. R., C. Baber & A. South (1996). Towards a definition and working model of stress and its effects on speech. *Speech Communication* 20(1): 3-12.

Neubauer, J., M. Edgerton & H. Herzel (2004). Nonlinear phenomena in contemporary vocal music. *Journal of Voice* 18(1): 1-12.

Nieto Caballero, O. (2008). *Voice transformations for extreme vocal effects*. Unublished. Universitat Pompeu Fabra. MA Thesis.

Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain.771-774

Nolan, F. (2012). The phonetic case for the involvement of native speakers in LADO. Paper presented at the *British Association of Academic Phoneticians Colloquium 2012*. Leeds, UK.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* 41: 1-16.

Ohala, J. J. (1996). Ethological theory and the expression of emotion in the voice. *Proceedings of The Fourth International Conference on Spoken Language Processing (ICSLP 96)*, Philadelphia, Pennsylvania.

Ohala, J. J. (2009). Signaling with the eyebrows – commentary on Huron, Dahl, and Johnson. *Empirical Musicology Review* 4: 101-102.

Petrushin, V. (1999). Emotion in speech: recognition and application to call centers. *Proceedings of Artificial Neural Networks in Engineering (ANNIE '99)*, St. Louis, Missouri.

Picard, R. W. (1997). *Affective Computing*. Cambridge: MIT Press.

Protopapas, A. & P. Lieberman (1997). Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America* 101: 2267-2277.

Riede, T., M. J. Owren & A. Clark Arcadi (2004). Nonlinear acoustics in the pant-hoot vocalizations of common chimpanzees (Pan Troglodytes): frequency jumps, subharmonics, biphonation, and deterministic chaos. *American Journal of Primatology* 64: 277-291.

Riede, T. & A. Stolle-Malorny (1999). The vocal change of a kitten with craniocerebellar trauma - a case study. *Bioacoustics* 10: 131-141.

Robb, M. P. & S. H. Saxman (1988). Acoustic observations in young children's non-cry vocalizations. *The Journal of the Acoustical Society of America* 83 1876-1882.

Roberts, L. (2008). *The Phonetics of Distress.* Unpublished. University of York.MSc Thesis.

Roberts, L. (2010). Real and acted responses of distress: an auditory and acoustic analysis of extreme stress and emotion. *Proceedings of the International Speech Communication Association (ISCA) Tutorial and Research Workshop (ITRW) on Experimental Linguistics (Exling 2010)*, Athens, Greece

Rose, P. (2009). Evaluation of disputed utterance evidence in the matter of David Bain''s retrial.[Online] Available at: http://rose-morrison.forensic-voice-comparison.net/documents/Rose%20(2009)%20Evaluation%20of%20disputed%20utterance%20evidence%20in%20the%20matter%20of%20David%20Bain%27s%20retrial%20(Report).pdf [Accessed 30 September 2012]

Rosengren, P. & A. Nilsson (1999). *Bumblebee Killer.* Unpublished. University of Karlskrona/Ronneby. MSc Thesis.

Rostolland, D. (1982). Phonetic structure of shouted voice. *Acta Acustica united with Acustica* 51(2): 80-89.

Ruiz, R., E. Absil, B. Harmegnies, C. Legros & D. Poch (1996). Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions. *Speech Communication* 20: 111-129.

Scherer, K., R., D. R. Ladd & K. E. A. Silverman (1984). Vocal cues to speaker affect: testing two models. *Journal of the Acoustical Society of America* 76(5): 1346-1356.

Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. *In* C. E. Izard (ed.) *Emotions in personality and psychopathology*. New York, Plenum. 495-529.

Scherer, K. R. (1981). Vocal indicators of stress. *In* J. K. Darby (ed.) *Speech Evaluation in Psychiatry*. New York, NY, Grune and Stratton. 171-187.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. *In* K. R. Scherer & P. Ekman (eds.) *Approaches to emotion*. Hillsdale, NJ, Erlbaum. 293–317.

Scherer, K. R. (1986). Vocal affect expression: a review and model for future research. *Psychological Bulletin* 99: 143-165.

Scherer, K. R. (1988). On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology* 7(2): 79-100.

Scherer, K. R. (1989). Vocal correlates of emotional arousal and affective disturbance. *In* H. Wagner and A. Manstead (eds.) *Handbook of Psychophysiology: Emotion and social behavior*. London, Wiley. 165-197.

Scherer, K. R. (2000). Psychological models of emotion. *In* J. C. Borad (ed). *The neuropsychology of emotion*. New York, Oxford University Press. 137-162.

Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication* 40: 227-256.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information* 44: 695-729.

Scherer, K. R., R. Banse, H.G. Wallbott & T. Goldbeck (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion* 15: 123-148.

Scherer, K. R., D. Grandjean, T. Johnstone, G. Klasmeyer, G. & T. Bänziger (2002). Acoustic correlates of task load and stress. *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)*, Denver, Colorado. 2017-2020.

Scherer, K. R., H. G. Wallbott & A. B. Summerfield (1986). *Experiencing Emotion: A Cross-Cultural Study.* Cambridge: Cambridge University Press.

Scherer, U., H. Helfrich & K. R. Scherer (1980). Internal push or external pull? Determinants of paralinguistic behavior. *Language: Social psychological perspectives* 279-282.

Schröder, M. (2001). Emotional speech synthesis: a review. *Proceedings of Eurospeech 2001*, Aalborg, Denmark.561-564.

Schröder, M., R. Cowie, E. Douglas-Cowie, M. Westerdijk & S. Gielen (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. *Proceedings of Eurospeech 2001* Aalborg, Denmark. 87-90.

Scripture, E. W. (1921). A study of emotions by speech transcription. *Vox* 31: 179-183.

Seifert, E. & J. Kollbrunner (2005). Stress and distress in non-organic voice disorders. *Swiss Medical Weekly* 135(27/28): 387.

Seyfarth, R. M. & D. L. Cheney (1986). Vocal development in vervet monkeys. *Animal Behaviour* 34(6): 1640-1658.

Sigmund, M. (2006). Introducing the database ExamStress for speech under stress. *Proceedings of the 7th Nordic Signal Processing Symposium (NORSIG 2006)*. Rejkjavik, Iceland. 290-293.

Simonov, P. V. & M. V. Frolov (1977). Analysis of the human voice as a method of controlling emotional state: achievements and goals. *Aviation, Space, and Environmental Medicine* 48: 23-25.

Sinclair, J. M. & R. M. Coulthard (1975). *Towards an Analysis of Discourse: The English used by teachers and pupils.* Oxford: Oxford University Press.

Skinner, E. R. (1935). A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness; and a determination of the pitch and force of the

subjective concepts of ordinary, soft, and loud tones. *Communications Monographs* 2(1): 81-137.

Sobin, C. & M. Alpert (1999). Emotion in speech: the acoustic attributes of fear, anger, sadness, and joy. *Journal of Psycholinguistic Research* 28: 347-365.

Spackman, M. P., B. L. Brown & S. Otto (2009). Do emotions have distinct vocal profiles? A study of idiographic patterns of expression. *Cognition and Emotion* 23: 1565-1588.

Stanislavsky, K. (1968). *Building a Character*. London: Methusen.

Streeter, L.A., R. M. Krauss, V. Geller, C. Olson & W. Apple (1977). Pitch changes during attempted deception. *Journal of Personality and Social Psychology* 35: 345-350.

Streeter, L.A., N.H. Macdonald, W. Apple, R. M. Krauss & K. M. Galati (1983). Acoustic and perceptual indicators of emotional stress. *Journal of the Acoustical Society of America* 73: 1354-1360.

Sundberg, J. & T. D. Rossing (1990). The science of singing voice. *Journal of the Acoustical Society of America* 87(1): 462-463.

Thomas, I. B. (1969). Perceived pitch of whispered vowels. *Journal of the Acoustical Society of America* 46(2B): 468-470.

Titze, I. R. (2008). Nonlinear source-filter coupling in phonation: theory. *Journal of the Acoustical Society of America* 123: 2733-2749.

Titze, I. R., T. Riede & P. Popolo (2008). Nonlinear source-filter coupling in phonation: vocal exercises. *Journal of the Acoustical Society of America* 123: 1902-1915.

Tokuda, I., T. Riede, J. Neubauer, M. J. Owren & H. Herzel (2002). Nonlinear analysis of irregular animal vocalizations. *Journal of the Acoustical Society of America* 111: 2908-2919.

Tolkmitt, F. J. & K. R. Scherer (1986). Effect of experimentally-induced stress on vocal parameters. *Journal of Experimental Psychology* 12: 302-313.

Traunmüller, H. & A. Eriksson (1997). A method of measuring formant frequencies at high fundamental frequencies. *Proceedings of EuroSpeech '97.* Rhodes, Greece.477-480

Traunmüller, H. & A. Eriksson (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America* 107: 3438-3451.

Vacher, M., J. F. Serignat, S. Chaillol, D. Istrate & V. Popescu (2006). Speech and sound use in a remote monitoring system for health care. *Text, Speech and Dialogue*. Berlin: Springer. 711-718.

Van Onze, B. (2005). "Grunten" sloopt de stem. *NederlandsDagblad.* [online] 28 June 2007. Available at: http://www.nd.nl/artikelen/2007/juni/28/-grunten-sloopt-de-stem. [Accessed 01 August 2012].

Velten, E. (1968). A laboratory task for induction of mood states. *Behaviour Research and Therapy* 6: 473-482.

Volodin, I. A. & E. V. Volodin (2003). Biphonation as a prominent feature of the Dhole *Cuos Alpinus* sounds. *Bioacoustics* 13: 105-120.

Webb, J. L. (2007). Promoting vocal health in the choral rehearsal. *Journal of Music Education* 93: 26-31.

Williams, C. E. & K. N. Stevens (1969). On determining the emotional state of pilots during flight: An exploratory study. *Aerospace medicine* 40: 1369-1372.

Williams, C. E. & K. N. Stevens (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America* 52: 1238-1250.

Wilting, J, E. J. Krahmer & M. G. J. Swerts (2006). Real vs. acted emotional speech. *Proccedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*. Pittsburgh, Pennsylvania. 805-808.

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law* 1(4): 792-816.

Yarmey, A. D. (2007). Earwitness descriptions and speaker identification. *International Journal of Speech Language and the Law* 8(1): 113-122.

Yarmey, A. D., A. L. Yarmey, M. J. Yarmey, & L. Parliament. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology* 15(3): 283-299.

Zuckerman, M., M. S. Lipets, J. H. Koivumaki, & R. Rosenthal (1975). Encoding and decoding nonverbal cues of emotion. *Journal of Personality and Social Psychology* 32(6): 1068-1076.

Court Cases

State of Florida -v- George Zimmerman (12 CF. 1083 A (Fla. 18th Cir. Ct. 2013))

The State Vs. Oscar Pistorius [2005] ZAGPPHC, High Courts – Gauteng, Pretoria (South Africa)

Court Case Reports

State of Florida -v- George Zimmerman, No 12-CF-1083 A. Court order excluding opinion of Mr. Owen and Dr. Reich [online]. June 22 2013. Available at: http://media.cmgdigital.com/shared/news/documents/2013/06/22/order_excluding_opinion_911_call.pdf [Accessed June 22 2013]