Effects of Manipulating Fundamental Frequency and Speech Rate on Synthetic Voice Recognition Performance and Perceived Speaker Identity, Sex, and Age

Georgina Elizabeth Gous

Thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

September 2017

ABSTRACT

Vocal fundamental frequency (F0) and speech rate provide the listener with important information relating to the identity, sex, and age of the speaker. Furthermore, it has also been demonstrated that manipulations in F0 or speech rate can lead to accentuation effects in voice memory. As a result, listeners appear to exaggerate the representation of a target voice in terms of F0 or speech rate, and mistakenly remember it as being higher or lower in F0, or faster or slower in speech rate, than the voice originally heard. The aim of this thesis was to understand the effect of manipulations/shifts in F0 or speech rate on voice matching performance and perceived speaker identity, sex, and age. Synthesised male and female voices speaking prescribed sentences were generated and shifted in either F0 and speech rate. In the first set of experiments (Experiments 2, 3, and 4), male and female listeners made judgements about the perceived identity, sex, or age of the speaker. In the second set of experiments (Experiment 5, 6, and 7) male and female listeners made target matching responses for voices presented with and without a delay, and with different spoken sentences. The results of Experiments 2, 3, and 4 indicated the following: (1) Shifts in either F0 or speech rate increased uncertainty about the identity of the speaker, though were more robust to shifts in speech rate than they were to shifts in F0. (2) Shifts in F0 also increased uncertainty about speaker sex, but shifts in speech rate did not. Male voices were accurately perceived as male irrespective of the direction of manipulation in F0. However, for female voices, decreasing F0 increased the uncertainty of speaker sex (i.e., the voices were more likely to be perceived as male rather than female). (3) Increasing either F0 or speech rate resulted in both male and female voices as sounding younger, whereas decreasing either F0 or speech rate lead to listeners perceiving the voices as sounding older. The results of Experiments 5, 6, and 7 indicated the following: (4) Shifts in

either F0 or speech rate did increase matching errors for the target voice, however, there was no evidence of an accentuation effect. Specifically, for voices shifted in F0, there was an increase in the selection of voices higher in F0 compared to voices lower in F0. For voices shifted in speech rate, there was an increase in the selection of voices faster in speech rate compared to voices slower in speech rate, but only for slow speech rate target voices. (5) Accentuation errors were no more likely to occur when the inter-stimulus interval was increased, or (6) when a different sentence was spoken in the sequential voice pair to the one previously spoken by the target voice.

The findings have theoretical and applied relevance. The work has provided a clearer understanding of how shifts in F0 or speech rate are likely to affect perceptions about the identity, sex, and age of the speaker than was possible to establish from previous studies. It has also contributed further to our understanding about the effect of shifts in F0 or speech rate on voice matching performance, and their importance in accurate recognition. This information might be insightful to the police and help to determine the accuracy of descriptions made about a voice and decisions made during a voice lineup, particularly if a suspect of a crime was likely to be disguising their voice.

ACKNOWLEDGMENTS

Where to start! This journey really has been an unforgettable one, and there are several people I would really like to thank for helping me along the way.

To my Director of Studies and good friend Dr Andrew Dunn who has given me such great support, guidance, and advice every step of the way. Whether it has been a personal issue or some thesis related disaster, you have always been there to calm me down and help me to make sense of it all again. You have taught me so much over the past four years and you have helped to shape me into the young academic I am today. There is so much that I have learnt from you and will take forward with me into the future. I can’t promise you that I won’t always write too much (because let’s face it, I always do!), but I will try my best! You have helped build my confidence and made me believe that I can succeed if I put my mind to it. I cannot thank you enough. To Professor Thom Baguley, thank you for all your advice and for being there to help me with some related statistical issue (and for bearing with me when I didn’t understand what you were telling me the first time around!). To Dr Paula Stacey, thank you for all your great advice you have given me and for helping me to formulate arguments and think more logically about things. And of course, for reading through my work and helping me shorten it down when necessary (which was quite often!). I feel so grateful to have had the guidance and support of such a fantastic supervisory team. Thank you all so much.

To my amazing Mum and Dad. I don’t even know where to begin. I really do owe you both so much. Thank you for making me into the person I am today and for loving me unconditionally, I am so incredibly grateful. I don’t know how I would have got through these past four years without the love and support you have both given me (actually, I’m not sure how I would have got through the past thirty years without you!). Anytime of the day or night you have been there

for me, dropped everything for me, you have literally done anything for me. You have listened to me laugh, you have listened to me cry, you have listened to me freak out when I thought that this couldn't be done (!), you have just always listened to me and been there for me, no matter what. You are both the most giving people and you really do mean everything to me. You have never stopped believing in me and you have supported me in every way you can. Financially, emotionally, I really am lost for words. I owe you both the world. You're just the best parents that anyone could ever ask for and I love you both so much. I know you will never ask for anything in return, because that's the kind of people you are. But I hope that one day I can give back what you have given to me. From the bottom of my heart Mum and Dad, thank you both so much.

To the love of my life, Andrew. By far the best part of this journey was meeting you along the way. You're my partner, my best friend, my soul mate, my world. I really do not know what I would do without you. You have supported me throughout all of this and have always been there with a big hug for me when I have just had enough. The ways you have cheered me up when I was lost in the midst of it all make me smile just thinking about them. You made me see the light on some very stressful days and have got me through the challenges of it all, every single time. You have given me the strength I need, and you have made me realise that I can achieve anything if I put my mind to it. You have showed me that I can do things I would never have imagined doing without you behind me, backing me all the way. Through the ups and the downs, I know we will always be there for each other, helping each other out and holding each other's hands along the way. Here's to the future and to all the wonderful memories we have to come and to look forward to. Thank you, Andrew, so much.

FINANCIAL SUPPORT

My PhD was supported by a three-year fully funded Vice Chancellor’s Scholarship awarded to me by Nottingham Trent University.

DECLARATION

This thesis comprises the candidate's own original work and has not, whether in the same or different form, been submitted to this or any other University for a degree. All experiments were designed and analysed by the candidate, and all testing was conducted by the candidate. Any publications that have occurred as a result of this thesis are the candidates own work.

Publications:

- Gous, G. E., Dunn, A. K., Baguley, T., & Stacey, P. (2017). An exploration of the accentuation effect: Errors in memory for voice fundamental frequency (F0) and speech rate. *Language, Cognition, and Neuroscience. 33,* 98-110.

Conference presentations/proceedings:

- Gous, G. E., Dunn, A. K., Baguley, T., & Stacey, P. (2017). *Errors in memory for disguised speech.* Presentation given as part of the Advancing Advocacy Conference, Nottingham Trent University.
- Gous, G. E., Dunn, A. K., Baguley, T., & Stacey, P. (2017). *The impact of variations in fundamental frequency (F0) and speech rate on voice recognition performance.* Presentation given as part of the Psychology Department External Seminar Series. Nottingham Trent University.
- Gous, G. E., Dunn, A. K., Baguley, T., & Stacey, P. (2016). *The impact of variations in fundamental frequency (F0) and speech rate on voice recognition performance.* Paper presented at EMUA conference, Loughborough University.
- Gous, G. E., Dunn, A. K., Baguley, T., & Stacey, P. (2016). *Voice recognition performance is reasonably robust to variations in speech rate, however, variations in*

fundamental frequency (F0) are more disruptive to memory. Paper presented at the Division of Psychology PhD Conference, Nottingham Trent University.

- Gous, G. E., Dunn, A. K., Baguley, T., & Stacey, P. (2015). *The effect of variations in fundamental frequency (F0) and speech rate on speaker recognition memory using a synthesised voice sample.* Poster presented at 32nd Annual BPS Cognitive Psychology Section Conference, University of Kent, Canterbury.
- Gous, G. E., Dunn, A. K., Baguley, T., & Stacey, P. (2014). *Earwitness memory: Factors that influence voice recognition accuracy across the lifespan.* Poster presented at 31st Annual BPS Cognitive Psychology Conference, Nottingham Trent University.
- Gous, G. E., Dunn, A. K., Baguley, T., & Stacey, P. (2014). *Factors that influence voice recognition accuracy and their forensic applications.* Presentation given as part of the Psychology Department Internal Seminar Series, Nottingham Trent University.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. OVERVIEW OF THESIS

1. Introduction

Manipulations in acoustic cues of the voice has been identified as important in determining accurate recognition of the voice and perceptions about characteristics of the speaker (Zhang, van de Weijer, & Cui, 2006). Understanding how people perform on tasks that involve these systems and the errors that can occur when doing so are likely to be of applied value and have important implications in the real-world. Studies of the extent to which listeners can judge a speaker's physical characteristics are common in the voice literature, in part because of the inherent interest on the topic (e.g., Hartman & Danhauer, 1976; Kreiman & Sidtis, 2013; Skuk & Schweinberger, 2013; Smith & Patterson, 2005). These studies have often identified that manipulations in acoustic cues of the voice, and particularly fundamental frequency (F0) and speech rate, can affect perceptual judgements for some of the characteristic information about the speaker (e.g., identity, sex, age, size, emotion etc.). Despite this however, the overall picture is still somewhat unclear with previous research has presented contradictory findings (e.g., Owren, Berkowitz, & Bachorowski, 2007; Gelfer & Bennett, 2013; Hillenbrand & Clark). There are also several methodological issues with the research that currently exist that limit the applicability of the findings. For example, some studies have used only one voice (e.g., Gaudrain, Li, Ban, & Patterson, 2009), others have used familiar speakers rather than unfamiliar speakers (e.g., Kuwabara & Takagi, 1991), and some have manipulated vowels or syllables rather than words or full sentences (e.g., Bennett & Montero-Diaz, 1982; Schwartz & Rine, 1968; Lavner, Gath, & Rosenhouse, 2000).

In contrast, few researchers have investigated the role of manipulations in acoustic cues of the voice and their impact on recognition performance for the voice. This is important because intra-speaker (i.e., within-speaker) variability in the voice exists. Intra-speaker variation is largely the result of the natural variation in vocal production. Speakers rarely pronounce given words or phrases in an identical way on different occasions, even if the second utterance is produced in close succession (Hollien, 1990). The same speaker can also sound different from time-to-time because of factors such as time of day, fatigue, mood and emotional state, changes in health, and intoxication (e.g., Nolan, 2005; Saslove & Yarmey, 1980). Intra-speaker variation can also occur when deliberately trying to disguise or modify the voice to sound different (e.g., to sound older, younger, or a different identity). As listeners, we are largely robust to these changes. Nevertheless, accurate recognition can be problematic and errors in memory can occur, particularly if the speaker is unfamiliar to the listener (e.g., Abberton & Foucin, 1978; Yarmey, Yaremy, & Yarmey, 2001; Ladefoged & Ladefoged, 1980; Zhang, 2012; Zhang & Tan, 2008). Of the few studies that do exist on this topic, research has found accentuation effects for voice memory where listeners mistakenly selected voices lower in F0 than low F0 target voices, and voices higher in F0 than high F0 target voices. For speech rate, listeners mistakenly selected voices slower in rate than slow rate target voices (Mullenix, Stern, Grounds, & Tessmer, 2010; Stern, Mullenix, Corneille, & Huart, 2007). The authors concluded that listeners rely on self-generated categorical information about the voice at the time of encoding to aid recognition when manipulations in F0 or speech rate are made (Mullenix et al., 2010; Stern et al., 2007. However, there are also several methodological issues with the research that currently exist that limit the applicability of the findings. For example, the researchers only used one male voice, and manipulations in F0 and speech rate fell outside the typical F0 and speech rate ranges of the English-speaking population. There may also be other factors that increase the likelihood that accentuation effects for voice memory will occur, including the time delay between hearing a voice and being asked to recognise this from a voice pair, and whether the voice speaks the same or a different sentence to the one previously heard. To date however, no research has explored these ideas further with voice stimuli.

Therefore, this thesis investigated the effect of manipulations in F0 or speech rate on voice recognition performance and perceptions about the speaker’s identity, sex, and age. The first three experiments reported here (Experiment’s 2, 3, and 4) investigated the extent to which manipulations in F0 or speech rate affect perceptions of a speaker’s identity, sex, and age for a set of unfamiliar male and female synthesised voices. The final three experiments (Experiment’s 5, 6, and 7) investigated the effect of manipulations in F0 or speech rate on recognition performance for a set of unfamiliar male and female synthesised voices. Overall, the findings suggested that the likelihood that a particular acoustic cue will affect perceptions about certain characteristics of the speaker is dependent on both the characteristic and the acoustic cue under investigation. Manipulations in F0 are likely to affect perceptions of the identity and age of the speaker. For female voices, decreasing F0 also increased the likelihood that the voices would be perceived as male. Manipulations in speech rate are unlikely to change perceptions of the identity or sex of the speaker. However, manipulations in speech rate do appear to affect perceptions of speaker age.

The findings also showed that listeners are susceptible to making errors for the voice when manipulations in F0 or speech rate are made. However, the findings are difficult to explain using the accentuation effect. Furthermore, for F0, listeners are no more likely to rely on self-generated categorical information about the voice at the time of encoding when the inter-stimulus interval is increased, or when a different sentence is spoken to the one that was previously heard.

1.1 Overview of Following Chapters

1.1.1 Chapter 2. The Human Voice: Producing a Voice and Controlling Its Sound

The purpose of Chapter 2 was to acquaint the reader with some relevant background information that is relevant for a fuller understanding of the thesis. The chapter provides an overview of the anatomy and physiology of human voice production, and the acoustic theory of speech production. It also includes an overview of the acoustic output of the speech signal. The chapter also explains some of the fundamental properties of speech, and provides a discussion of how speakers can deliberately manipulate their vocal apparatus to change different aspects of the sounds they produce.

1.1.2 Chapter 3. Literature Review: Speaker Perception and Recognition Memory

Chapter 3 places the thesis within the wider context of the existing literature. It begins by discussing differences that exist in fundamental frequency (F0) and speech rate of speakers of different identities, male and female speakers, and speakers of different ages. This section also reviews evidence that has considered manipulations in F0 or speech rate and how they affect perceptual judgements about the identity, sex, and age of the speaker. However, it comes to light that, despite this, the overall picture is still somewhat unclear. The review also highlights several methodological issues with the studies that currently exist on these issues, which in turn, provides a rationale for investigating this topic further.

Chapter 3 then moves on to discuss the impact of manipulations in F0 or speech rate on recognition performance for the voice. It begins by reviewing evidence that has considered the effect of placing stimuli into distinct categories, by demonstrating how memory has been found to often reflect typical representations of a stimulus rather than specific features of those learned items, also known as the accentuation effect. It is apparent that very few researchers have considered categorisation or accentuation effects in relation to voices, and highlights several methodological constraints with the studies that do currently exist on this issue. Chapter 3 also considers whether listeners become increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition when other factors are introduced. It discusses the time course of echoic memory, and reviews research that has found people to be more accurate in a task that uses shorter intervals between presentations of the stimuli. It also discusses how memory for the voice may be somewhat easier if the sentence spoken is the same throughout the duration of the task. In discussing the existing research in detail, several gaps in knowledge emerge. On the basis of this, the following research questions were formulated:

- **(1):** Do manipulations in fundamental frequency (F0) or speech rate affect perceptual judgments about the paralinguistic characteristics of the speaker, and if so, how do they change? Specifically, how do manipulations in F0 or speech rate affect perceptions of;

  a) speaker identity?

  b) speaker sex?

  c) speaker age?

- **(2):** Do manipulations in fundamental frequency (F0) or speech rate affect recognition performance for voices, and if so, can the findings be explained using the accentuation effect?

- **(3):** Do listeners become increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition when;

a) F0 is increased and decreased, and the inter-stimulus interval between presentation of the target voice and the sequential voice pair is increased?

b) F0 is increased and decreased, and a different sentence is spoken in the sequential voice pair to the one previously spoken by the target voice?

1.1.3 Chapter 4. Stimuli Development

Chapter 4 outlines the methods used to develop the voice stimuli (3 male and 3 female) in this thesis. It begins by explaining how the voices were manipulated and the measurements that were calculated for both F0 and speech rate. It then moves on to discuss several factors (i.e., speaker familiarity, ethnicity and accent of voices, emotional stress and arousal, and voice sample durations) that have been found to affect the performance of listeners in speaker perception and recognition tasks, and explains how these have been controlled for during the experiments. The chapter also reports five experiments that were carried out to obtain information about the voices (i.e., perceived similarity of the voices, perceived identity of the voices, naturalness ratings of the voices), and to ensure that the stimuli used for the experiments were appropriate.

1.1.4 Chapter 5. Speaker Perception: Identity

Chapter 5 reports on Experiment 2 which investigated whether manipulations in F0 or speech rate affect perceptions about the identity of the speaker. A 2AFC perceptual discrimination paradigm was used in which listeners were presented with within voice pairs whereby one of the six original voices was paired with a manipulated version of that voice (i.e., increased or decreased in F0 or speech rate). The listeners task was to decide whether the pair of voices presented were the same identity or a different identity. The results suggested that whilst greater manipulations in both F0 or speech rate increased uncertainty about the identity of the speaker, listeners were more robust to changes in speech rate than they were to changes in F0. It was concluded that F0 is more directly related to speaker identity than speech rate, and that listeners rely on F0 more than they do on speech rate when making decisions about the identity of the speaker.

1.1.5 Chapter 6. Speaker Perception: Sex

Chapter 6 reports on Experiment 3 which investigated whether manipulations in F0 or speech rate affect perceptions about the sex of the speaker. The listeners were presented with one of the six original voices or manipulated versions of the voices (i.e., increased or decreased in F0 or speech rate) and asked to decide whether the voice they heard was male or female. The results suggested that manipulations in F0 were more likely to increase uncertainty about speaker sex than manipulations in speech rate. Although voices that were decreased in speech rate did increase the uncertainty of speaker sex, overall listeners were accurate at determining speaker sex when voices were manipulated in speech rate. Whilst male voices were accurately perceived as male irrespective of the direction of manipulation in F0, for female voices, decreasing F0 increased the uncertainty of speaker sex (i.e., voices were more likely to be perceived as male rather than female). It was concluded that F0 is more directly related to speaker sex than speech rate, and that listeners rely on F0 more than they do on speech rate when making decisions about the sex of the speaker.

1.1.6 Chapter 7. Speaker Perception: Age

Chapter 7 reports on Experiment 4 which investigated whether manipulations in F0 or speech rate affect perceptual judgements about the sex of the speaker. The listeners were presented with one of the six original voices or manipulated versions of the voices (i.e., increased or decreased in F0 or speech rate) and asked to freely estimate the age of the speaker.

The results suggested that manipulations in both F0 and speech rate were likely to affect perceptions of speaker age. For both male and female voices, increasing F0 or speech rate lead to listeners perceiving the voices as sounding younger, whereas decreasing F0 or speech rate lead to listeners perceiving the voices as sounding older. However, some discrepancy appeared to exist between listeners expectations about speakers of different ages and the vocal characteristics that actually exist. It was concluded that both F0 and speech rate are important cues for estimating speaker age.

1.1.7 Chapter 8. Recognition Memory: An Exploration of the Accentuation Effect

Chapter 8 reports on Experiment 5a and 5b which investigated whether manipulations in F0 (Experiment 5a) and speech rate (Experiment 5b) affect recognition performance for the voice, and if so, whether the findings are attributable to the accentuation effect. Using a 2AFC procedure, the listeners were presented with a target voice before being presented with a sequential voice pair that included the previously heard target voice and a manipulated version of the voice. A 1-second inter-stimulus interval was used between presentation of the target voice and the sequential voice pair. The listeners task was to decide which voice in the sequential voice pair (voice 1 or voice 2) was the voice they had previously heard. The results showed that manipulations in F0 and speech rate increased recognition errors. For F0, there was an increase in the selection of voices higher in F0 compared to voices lower in F0 for high, moderate, and low F0 target voices. For speech rate, there was an increase in the selection of voices faster in speech rate compared to voices slower in speech rate for slow speech rate target voices. However, there was no difference in the selection of voices faster and slower in speech rate for fast and moderate speech rate target voices. It was concluded that the findings were difficult to explain using the accentuation effect.

1.1.8 Chapter 9. Recognition Memory: Increasing the Inter-Stimulus Interval

Chapter 9 reports on Experiment 6 which investigated whether manipulations in F0 affect recognition performance for the voice when the inter-stimulus interval between presentation of the target voice and the sequential voice pair was increased to 5-seconds, and if so, whether the findings were attributable to the accentuation effect. This experiment was designed to push the target voice out of the range, or at least to the very limits, of sensory memory. Using a 2AFC procedure, the listeners were presented with a target voice before being presented with a sequential voice pair that included the previously heard target voice and a manipulated version of the voice. A 5-second inter-stimulus interval was used between presentation of the target voice and the sequential voice pair. The listeners task was to decide which voice in the sequential voice pair (voice 1 or voice 2) was the voice they had previously heard. Overall the pattern of results observed in Experiment 6 were largely similar to those observed in Experiment 5a. It was concluded that listeners were no more likely to rely on self-generated categorical information about the voice at the time of encoding to aid recognition when the inter-stimulus interval is increased.

1.1.9 Chapter 10. Recognition Memory: Changing the Spoken Message

Chapter 10 reports on Experiment 7 which investigated whether manipulations in F0 affect recognition performance for the voice when a different sentence was used in the sequential voice pair to the one previously spoken by the target voice, and if so, whether the findings were attributable to the accentuation effect. Using a 2AFC procedure, the listeners were presented with a target voice before being presented with a sequential voice pair that included the previously heard target voice and a manipulated version of the voice. The voices in the sequential voice pair spoke a different sentence to the previously heard target voice. A 1-second inter-stimulus interval was used between presentation of the target voice and the sequential voice pair. The listeners task was to decide which voice in the sequential voice pair (voice 1 or voice 2) was the voice they had previously heard. Overall, the pattern of results observed in Experiment 7 were largely similar to those observed in Experiment 5a and 6. It was concluded that listeners were no more likely to rely on self-generated categorical information about the voice at the time of encoding to aid recognition when a different sentence is spoken to the one previously spoken by the target voice.

Chapter 10 also reports on Experiment 8 which set out to determine whether the amount of matching errors made overall were different for the three recognition memory experiments (Experiment 5a, Experiment 6, and Experiment 7) when manipulations in F0 were made. Experiment 8 explored whether accurate matching decisions rely on high quality representations temporarily stored in echoic memory (Experiment 5a, Chapter 8) and whether the ability to make accurate decisions diminishes as the inter-stimulus interval increases (Experiment 6, Chapter 9). Experiment 8 also explored whether listeners were more accurate at matching voices when the content of the sentences in the sequential voice pair was the same as the sentence spoken by the target voice (Experiment 5a, Chapter 8) compared to when the content of the sentence in the sequential voice pair was different to the sentence spoken by the target voice (Experiment 7, Chapter 10). The findings suggested that listeners made fewer matching errors overall when the same sentence was used in the sequential voice pair as the previously heard target voice compared to when a different sentence was used in the sequential voice pair to the previously heard target voice. It was concluded that listeners use both similarities in elements of the spoken message and F0 of the voice to help aid the recognition process. There was no difference in errors made overall when a 1-second inter-stimulus interval was used and when a 5-second inter-stimulus interval between presentation of the target voice and the sequential voice pair. It was concluded that recognition performance for voices may not be directly dependent on the time course of echoic memory. Rather, the findings suggested that the representation of vocal stimuli in auditory memory may be retained for periods longer than 3 to 5 seconds, and may leave a stronger trace in memory than non-vocal auditory stimuli.

1.1.10 Chapter 11. Summary of Findings and General Discussion

Chapter 11 summarises the aims, main findings and conclusions from the experiments reported in this thesis. It also discusses the potential applied implications of these findings, limitations of the stimuli set, and suggests some outstanding research questions and possible future directions for research.

CHAPTER 2. THE HUMAN VOICE:

PRODUCING A VOICE AND CONTROLLING ITS SOUND

2. Introduction

The purpose of this chapter is to acquaint the reader with some background information that is relevant for a full understanding of the thesis. It provides an overview of the anatomy and physiology of human voice production, the acoustic theory of speech production, and an overview of the acoustic output of the speech signal. The chapter also explains some of the fundamental properties of speech, and discusses how speakers can manipulate their vocal apparatus to change different aspects of the sounds they produce.

2.1 The Human Voice

The human voice is the carrier of speech and is the most important sound of our auditory environment (Belin, Fecteau & Bedard, 2004). As humans, we probably spend more time everyday listening to voices than to any other sound, and our ability to analyse and categorise information contained in voices plays a key role in human social interactions (Belin, Fecteau & Bedard, 2004). It is generally accepted that the voice evolved as an aid to survival in the environment and as a means of communication (Benninger, 2010). Physiologically speaking, vocal production is similar across all primates (Ghazanfar & Rendall, 2008). The notable difference between humans and other primates however is their ability to produce and understand speech (Belin et al., 2004). For the purposes of this thesis, it is important to make a distinction between several concepts that will be used throughout the work. The voice will refer to the sound produced by a person’s vocal equipment and uttered through the mouth as speech (Traunmuller & Eriksson, 2000) (this will be discussed further in Section 2.1.1). Speech will refer to the vocalised form of human communication that conveys information between a speaker and a listener on several layers (Laver, 1991) (which will be discussed fully in Section 2.2.1). Finally, the speaker will refer to a person who produces speech. In this way, a speaker is an individual and speech is an artefact of a process that goes on within that speaker (Laver, 1991).

2.1.1 Anatomy of the Vocal System

The main components important in the production of speech are shown in Figure 2.1.

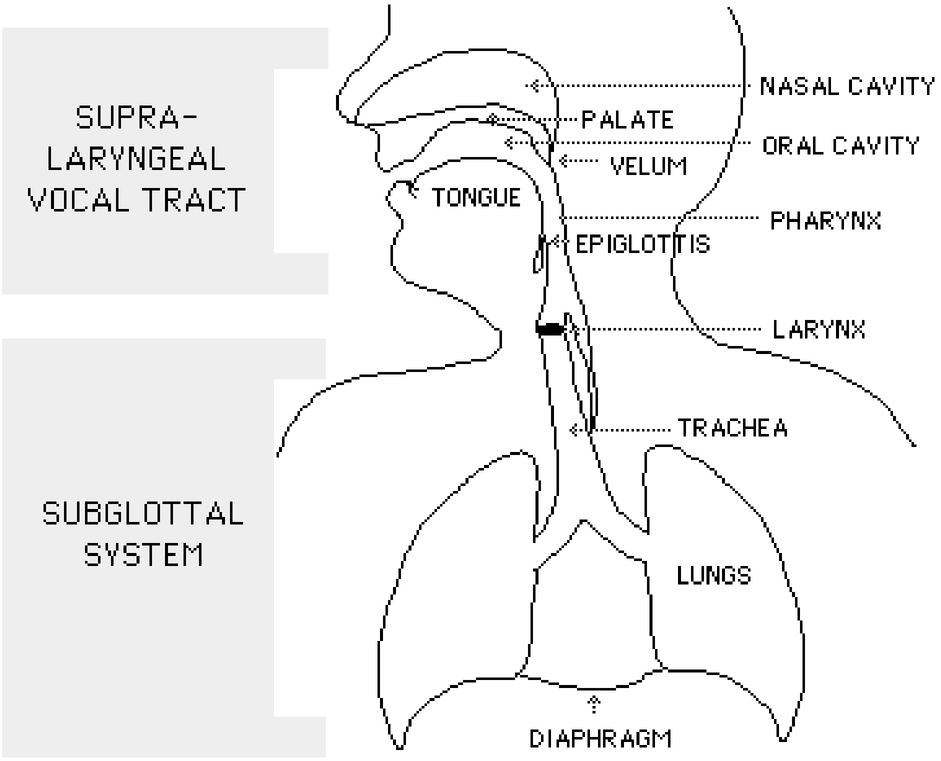


Figure 2.1: The human speech production system. Reprinted from Rubin, P., & Vatikiotis-Bateson, E. (1998). Measuring and modelling speech production. In Hopp, S. L., Owren, M. J., & Evans, C. S (Ed.), Animal Acoustic Communication: Sound Analysis and Research Methods (pp. 251-282). New York: Springer-Verlag.

The *subglottal system* refers to all those features of the vocal system situated below the larynx (Ardran & Kemp, 1996). This includes the *diaphragm* which is the primary muscle used in the process of inspiration or inhalation, the *lungs* which allow the body to take in oxygen from the air, and the *trachea*, commonly known as the windpipe, which is a tube that travels from the larynx to the lungs (Ardran & Kemp, 1996).

Situated above the subglottal system is the *larynx*, or the voice box, which is an organ that contains the vocal folds, or vocal cords. The *glottis* is the gap between the vocal folds and the *epiglottis* is the flap of elastic cartilage tissue attached to the entrance of the larynx (Ardran & Kemp, 1996).

The air passages above the larynx are known collectively as the *supra-laryngeal vocal tract.* The vocal tract can be divided into the *oral cavity* and the *nasal cavity* (Hopp, Owren, & Evans, 1997. The *oral cavity* includes the lips, cheeks, teeth, tongue, soft palate (velum) and hard palate (roof of the mouth) (Hopp et al., 1997). The *nasal cavity* is a large air filled space above and behind the nose (Hopp et al., 1997). The *pharynx* is a tube which begins just above the larynx. At the top end, the pharynx is divided into the oral and nasal cavities (Hopp et al., 1997. The cavities of the supra-larygneal vocal tract are called resonating cavities (Ardran & Kemp, 1996). Resonation is the process by which phonated sounds are enhanced in intensity by the air-filled cavities through which it passes (McKinney, 1994).

2.1.2 Acoustic Theory of Speech Production

The prominent acoustic theory of speech production is the source-filter theory (Fant, 1960), and describes speech production as a two-stage process (Fant, 1960). In this way, acoustic speech output is considered to be the combination of a *source* of sound energy (i.e., the larynx) modulated by a filter function determined by the shape of the supra-laryngeal vocal tract (Yehia, Rubin & Vatikiotis-Bateson, 1998).

2.1.2.1 The Source

The source of the sound energy (i.e., the larynx) is shown in Figure 2.2.

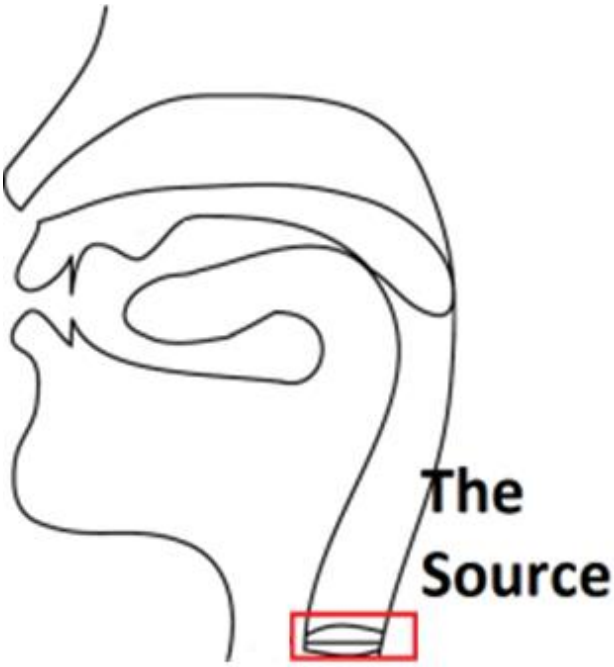


Figure 2.2: Diagram depicting the source of the speech signal (in red). This is composed of the larynx, and specifically the vocal folds.

Voice production essentially begins with respiration (breathing). Air is inhaled as the diaphragm lowers causing the volume of the lungs to expand as air rushes in to fill this space (Harrington & Cassidy, 2012). As we exhale, the muscles of the rib cage lower and the diaphragm raises, essentially squeezing the air out (Lapena & Calaquian, 2004). This action supplies the air stream responsible for the production of the speech signal, as well as for breathing. The air travels up the trachea where it reaches the larynx (Rao & Koolagudi, 2012). Air is then pushed past the vocal folds in the larynx. If the air is pushed past the vocal folds with sufficient pressure, they begin to vibrate and phonation occurs. If however the vocal folds in the larynx do not vibrate, speech is produced as a whisper (Clark & Yallop, 1995). The air flow is chopped into a sequence of quasi-periodic pulses through the vibration of the vocal folds (Harrington & Cassidy, 2012) (which will be discussed fully in Section 2.1.3.2). The rate, or frequency, at which the puffs of air exit the larynx is known as the fundamental frequency

(F0) of the laryngeal source and contributes to the perceived pitch of the produced sound (Rubin & Vatikiotis-Bateson, 1998). The rate at which the vocal folds open and close during phonation can be varied in several ways and is determined by the tension of the laryngeal muscles and the air pressure generated by the lungs (Rubin & Vatikiotis-Bateson, 1998).

2.1.2.2 The Filter

Having passed through the larynx, the speech signal then undergoes further changes as it makes its way up towards the mouth. The sound wave produced by the vocal folds are too weak to be recognised as a voice, and so must be amplified for listener audibility (Clark & Yallop, 1995. For this signal to be not only audible, but also structured in a way that it can transmit linguistic information (this will be discussed more fully in Section 2.2.1), parts of the vocal tract must be controlled and co-ordinated so that the acoustic variations in the signal conform to the language being spoken (Clark & Yallop, 1995). The pharynx, oral, and nasal cavities of the supra-laryngeal tract are resonators that act as acoustic filters to the original source of sound (Harrington & Cassidy, 2012). Figure 2.3 shows the location of these filters in the human vocal tract.

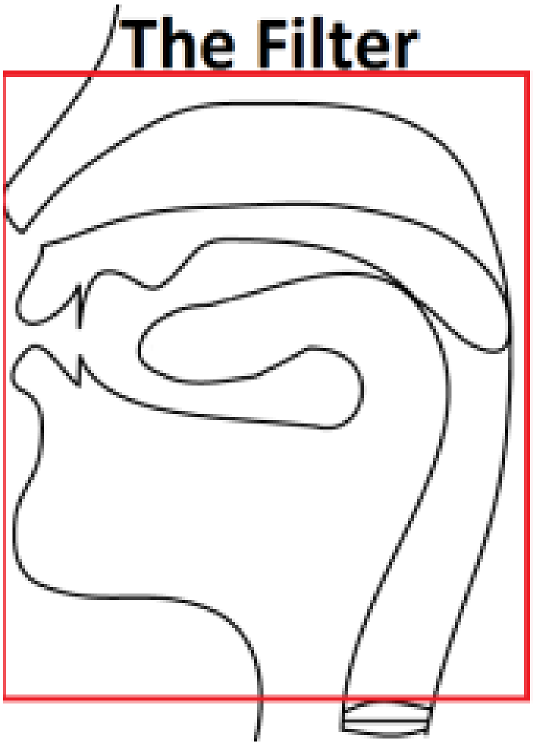


Figure 2.3: Diagram depicting the location of the voice filters in the human vocal tract (in red). This is composed of the supra-laryngeal vocal tract, and specifically the oral and nasal cavities.

The cavities in the supra-laryngeal vocal tract acts as acoustic filters by adjusting the relative intensities of the frequency components of the sound (Rendall, Vokey, & Nemeth, 2007; Xu, Homae, Hashimoto & Hagiwara, 2013). Energy at frequencies that coincide with the natural resonance frequencies of these airways passes easily and with greater amplitude, while energy at other frequencies is attenuated by being absorbed by the vocal tract walls (Ghazanfar & Rendall, 2008). As we speak, the cavities of the supra-laryngeal vocal tract are constantly changing shape, which determines the frequencies that are accentuated and the frequencies that are attenuated (Ghazanfar & Rendall, 2008). Humans possess a large and complex set of muscles that produce changes in the shape of the vocal tract, known as articulators (Garnier, Wolfe, Henrich, & Smith, 2008). Changing the shape of the vocal tract leads to changes in the resonances of the vocal tract, and thereby amplifying the sounds (Garnier et al., 2008). Articulators can be subdivided into those that are active (i.e., those that move), including the lips, tongue and velum, and those that are passive (i.e., those that do not move), such as the teeth and hard palate. It is the movement of these active articulators that are used to form recognisable sounds and words (Garnier et al., 2008). It is these variations in pressure and flow produce in the acoustic signal that we hear when listening to speech (McGowan, 1994).

2.1.3 The Acoustic Output

2.1.3.1 The Sound Wave

The sound wave signal produced by the vocal tract is a pattern of disturbance caused by the movement of energy travelling through a medium as it propagates away from the source of the sound (Taylor, Reby & McComb, 2011). In terms of the sound produced by the human vocal tract, the medium that the sound wave travels through is air. The resulting vibrations produced by the human vocal tract radiates from the mouth and nose into the environment, disturbing the surrounding particles in the air (Taylor et al., 2011). This in turn disturbs those particles next to them, and so on, resulting in changes in the pressure of the surrounding air. This pattern of disturbance travels steadily away from its source and creates an outward movement in a wave pattern. Sound waves are longitudinal waves, which means the direction of vibrations in the air is the same as the direction of travel of the wave (Plack, 2013).

Sound is a waveform compromising of amplitude/intensity and frequency that vary as a function of time Taylor et al., 2011). These waves can be graphed on a Cartesian coordinate plane, where the waveform is a plot of amplitude against time (Taylor et al., 2011). Figure 2.4 illustrates a visual representation of a sound wave.

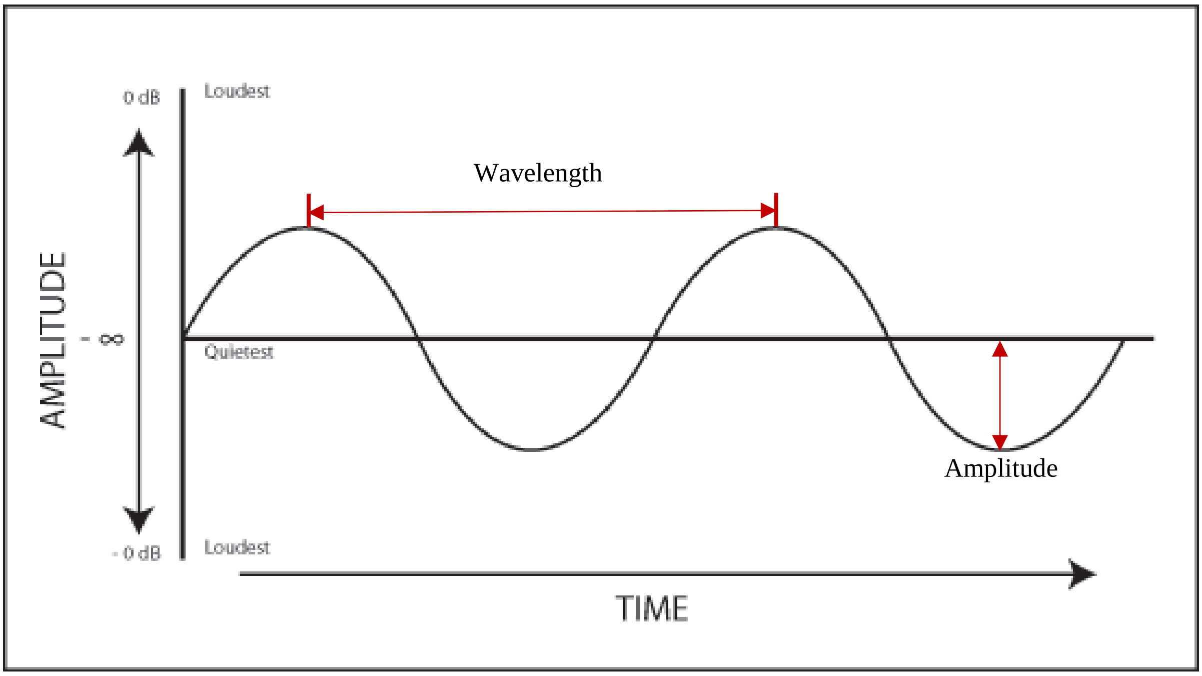


Figure 2.4: The main characteristics of a sound wave. Adapted from http://www.studio-diy.net/category/recording/.

Time is plotted along the *x* axis and represents how the pressure of the air and frequency of the sound changes over time (Plack, 2013). Time is often measured in milliseconds (ms). *Amplitude* is plotted on the *y* axis and is usually measured in decibels (dB) (Plack, 2013). The amplitude of a wave is its maximum disturbance from its undisturbed position. In the case of sound waves, it is the extent of the maximum variation in air pressure from normal atmospheric pressure (the central horizontal line, or baseline, in Figure 2.4) (Taylor et al., 2011). The sound wave behaves as an alternating current, meaning that the amplitude changes from areas of high pressure (compressions) to areas of low pressure (rarefactions), and then back again (Plack, 2013). Humans perceive amplitude as loudness. The further the wave is from the central line, the higher the amplitude, and the louder the wave will sound. However, when the wave is closer to the central line, the amplitude is lower and the sound will be quieter. The *wavelength* of a wave is the physical distance between the point on one wave and the same point on the adjacent wave (i.e., the distance between two pressure peaks or two pressure troughs) (Garnier et al., 2008). The *period* of the waveform is the time taken for a complete pattern of repetition (or one cycle). It is measured in seconds (s) and its fractions (milliseconds, nanoseconds, etc.) (Garnier et al., 2008). The *frequency* of a wave is the speed of vibration, and is measured as the number of wave cycles that occur in one second. The most commonly used unit of measurement for frequency is cycles per second, or Hertz (Hz) (Plack, 2013).

As previously noted (in Section 2.1.2.1), the sensation of the frequency of a sound is closely related to the perception of pitch. A high pitch sound corresponds to a higher frequency sound, whereas a low pitch sound corresponds to a lower frequency sound. Higher frequency waves tend to be shorter and more compressed, whereas lower frequency waves tend to be longer and less compressed (Plack, 2013). Accordingly, higher frequency waves produce more cycles per second than lower frequency waves (Plack, 2013). This is illustrated in panel A and panel B of Figure 5.

2.1.3.2 Complex Sound Waves

So far, a sound wave has only been discussed in terms of one frequency component being present. These are known as simple waves and resemble sine waves when plotted (Ladfoged, 1962). Simple waves are heard as pure tones (Plack, 2013). The addition of simple waves of different frequencies result in a complex wave. Speech is an example of a complex wave and is therefore composed of more than one frequency (pure tone). The repetition rate of a complex tone is known as the fundamental frequency (F0) and is measured in Hz. Figure 2.5 provides an illustration of two pure tones (panel A and panel B) combined to form a complex wave (panel C).

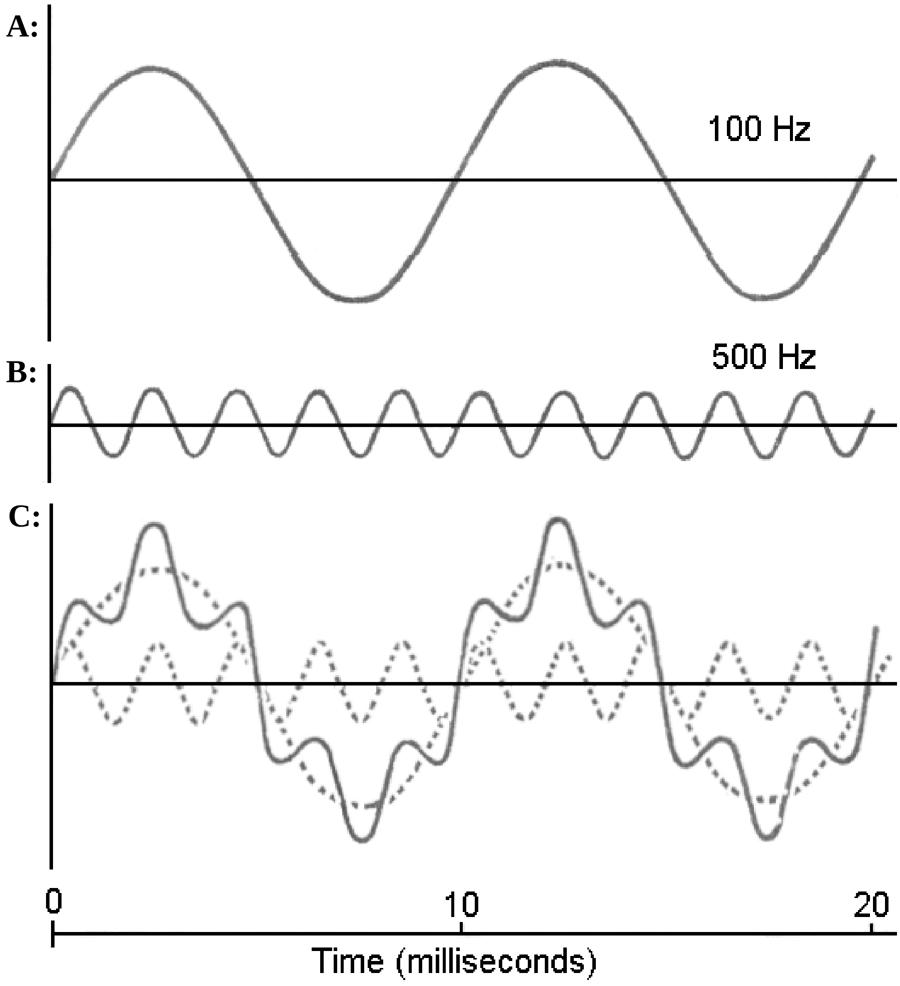


Figure 2.5: An example of adding together two simple waves to form a more complex wave (panel C). Retrieved from Ladfoged (1962). It should also be noted that the wave in panel B is half the amplitude of the wave in panel A.

Figure 2.5 to shows how the summation of two simple waves (panel A and panel B) can form a more complex wave (solid line in panel C). Panel A shows a simple sound wave of 100 Hz, and panel B shows one at 500 Hz. Panel C shows the resulting wave (solid line) when the two simple waves (dashed lines) are summed together. Compared to the waves in panel A and B, the wave in panel C has a more complex pattern (Ladfoged, 1962). The fundamental shape of the wave is a representation of the intensity of energy that is produced when these simple waves are overlapped and summed up to form a complex wave (Ladfoged, 1962).

Complex sound waves can be both periodic and aperiodic in nature (Plack, 2013). Periodic waves refer to those where the resulting periods are identical and evenly timed to its adjacent periods (Plack, 2013). The complex wave (panel C) shown in Figure 2.5 is an example of a periodic waveform. Aperiodic waves however are those where successive disturbances are not identical and evenly spaced in time to its adjacent periods (Plack, 2013).

Speech is characterised by both the presence and absence of periodic and aperiodic waves. In this way, the speech signal is quasi-periodic, meaning almost periodic, or not wholly repeating (Remy & McComb, 2003). Being an almost periodic function, means that any one period is virtually identical to its adjacent periods, however it is not necessarily similar to periods further away in time (Plack, 2013). In speech, the degree of periodicity in the vocal signal is determined by the vibration of the vocal folds (Remy & McComb, 2003). Periodic sound waves are the result of regular excitation of the vocal folds (Plack, 2013). In other words, all periodic speech sounds are phonated. Vowel sounds have periodic waveforms as they are produced by a voiced source (Crystal, 2006). Aperiodic waves are the result of unvoiced speech (Taylor et al., 2011). Sources of unvoiced speech include a brief pulse of excitation caused by a rapid change in oral air pressure, and turbulence noise generated as air flows rapidly through an open, non-vibrating glottis (i.e., aspiration), or a narrow constriction of the supra-laryngeal vocal tract (i.e., frication) (Diehl, 2008). Such sources contain no periodic component, and consequently form irregular patterns in the sound wave (Diehl, 2008). A number of consonants, such as fricatives (e.g., /f/ and /s/) and stop consonants (e.g., /p/ and /t/), have aperiodic waveforms as they are associated with a rapid reduction in oral air pressure at the moment of vocal tract opening (Fant, 1973). Both periodic and aperiodic sources of sounds have an energy level sufficient to evoke a response from the resonances in the vocal tract and generate highly audible sounds (Diehl, 2008). Figure 2.6 depicts a complex quasi-periodic waveform of a person speaking the utterance *“on our website”*.

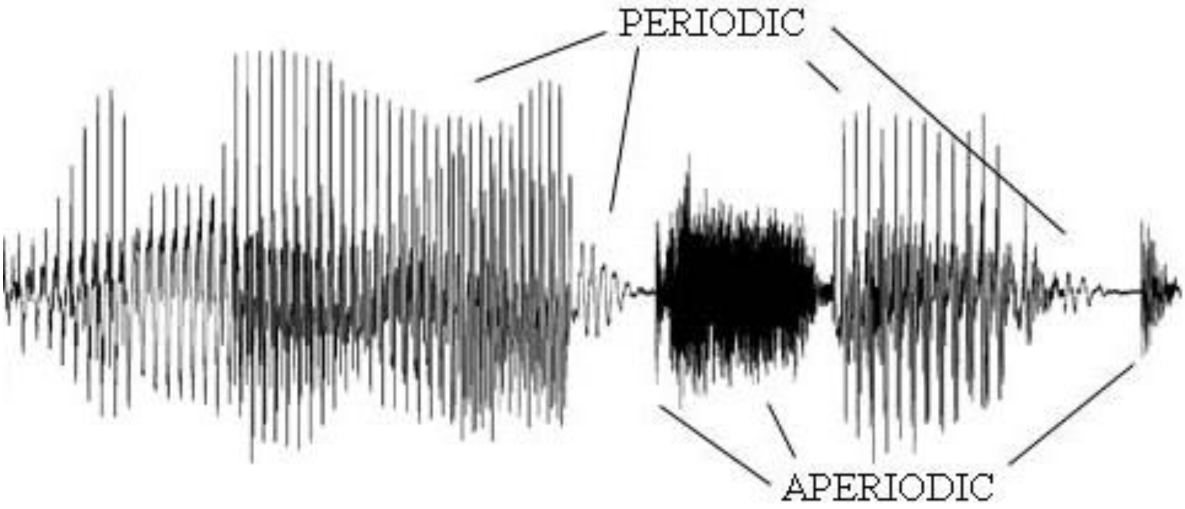


Figure 2.6: Complex quasi-periodic waveform of the utterance "on our website". Retrieved from. http://swphonetics.com/praat/tutorials/understanding-waveforms/speech-waveforms/.

2.1.3.3 Fundamental Frequency (F0) and Harmonics

As previously noted, both complex periodic and quasi-periodic sound waves are made up of a series of pure tones at specific frequencies. The frequency components of complex waves are called harmonics (Plack, 2013). The first component in a harmonic series is the fundamental frequency (F0). F0 refers to the lowest frequency, and thus the slowest repeating component of the period. F0 is also the main acoustical cue that determines the perceived pitch of speech (Reby & McComb, 2001). A higher pitch has a higher F0, whereas a lower pitch has a lower F0. All remaining harmonics are integer multiples of the F0, meaning that the harmonic spacing equals the F0 of vocal fold vibrations (Reby & McComb, 2011). Successive harmonics can be found by repeatedly adding F0, and vibrate at 2, 3, 4 times (etc.) as fast as F0. For example, if the F0 is 100 Hz, the second harmonic (H2) will be 200 Hz, the third harmonic (H3) will be 300 Hz, the fourth harmonic (H4) will be 400 Hz, and so on. The more complex the wave, the more frequency components there are (Plack, 2013). Speech is very complex and vocal fold vibration produces many harmonics above an F0 that range all the way up to 5000

Hz in the adult human vocal tract (Plack, 2013). These harmonics decrease in amplitude as the frequency increases (Plack, 2013). What we actually hear as speech therefore are many different pure tones summed together to form a more complex tone/waveform. However, the human ear does not typically perceive harmonics as separate frequencies, rather it is heard as one sound.

2.1.3.4 The Spectrogram

An alternative way to view sound is using a spectrogram. A spectrogram is a visual representation of an acoustic signal, where time is plotted on the *x* axis and frequency is plotted on the *y* axis (Plack, 2013). Using a mathematical technique known as Fourier analysis, the complex sound wave can be separated into the frequencies and amplitudes of its component sine waves (Jansen & Niyogi, 2006). The amount of energy (amplitude) at a particular combination of frequency and time is displayed as variations in greyscale darkness (with white meaning no energy and black meaning a high degree of energy) (Liberman, Latiman, Reindenberg & Gannon, 1992).

Narrow band spectrograms can be created to identify both the F0 and harmonics of speech (Liberman et al., 1992). Narrow band spectrograms are created using very fine, high resolution frequency analysis (Jansen & Niyogi, 2006). This analysis is fine-grained enough to reveal the rich harmonic content of voiced speech, but smears together adjacent moments in time (Jansen & Niyogi, 2006). Figure 2.7 illustrates a narrow band spectrogram of the spoken word *“heard”.*

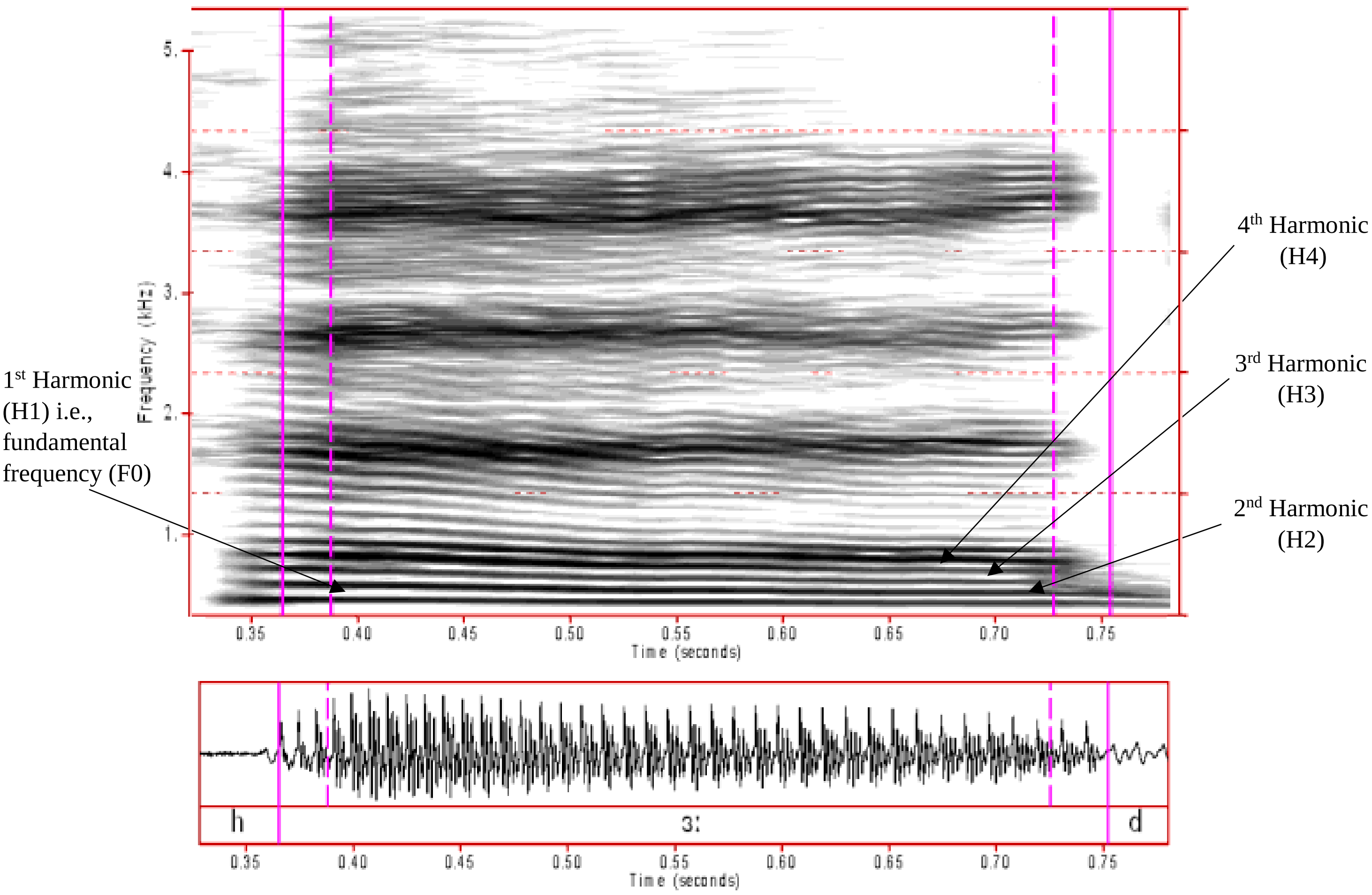


Figure 2.7: Narrow band spectrogram and sound wave for the spoken word *"heard"*, using band pass filters with a band width of approximately 45 Hz. Adapted from http://clas.mq.edu.au/speech/acoustics/frequency/spectral/html. Darker areas indicate greatest energy (amplitude). The vertical pink lines illustrate the beginning and end of a speech sound.

The narrow band spectrogram in Figure 2.7 reveals horizontal striations, with each band representing the different harmonics of the spectrum (Plack, 2013). The first harmonic (or F0) is depicted by the lowest striation, with all other consequtive striations revealing a different harmonic (Plack, 2013). The first four harmonics have been highlighted (e.g., H1, H2, H3, H4). Because some harmonics are stronger than others at any given time (owing to resonances of the vocal tract), these are also apparent. The sound wave for the utterance is also present, allowing a comparison to be made between the two visual representations of the sound's source (Plack, 2013). The vertical pink lines divide both the spectrogram and the sound wave into moments in time. The individual speech sounds of the utterance are also emphasised.

Specifically, the uttterance can be broken down into three speech sounds; /h/, /ear/ (phonetic symbol /ɜː/), and /d/.

2.1.3.5 Formants

So far, speech has been discussed in terms of its component frequencies (i.e., F0 and harmonics), thus describing the resulting sound that is made by the source (i.e., the vocal folds in the larynx). The vocal tract acts as a filter to the source, and speech sounds are characterised by a number of different articulations of the supra-laryngeal vocal tract (Ghazanfar & Rendall, 2008; Latinus & Belin, 2011; Reby & McComb, 2003). Different vocal tract configurations yield different filters, and these determine what component frequencies resonate for a particular speech sound (Ghazanfar & Rendall, 2008; Latinus & Belin, 2011; Reby & McComb, 2003). The resulting peaks in frequencies are called formants (Fant, 1960). Formants represent speech sounds, emphasising certain frequencies at higher amplitudes (Fant, 1960). The formant with the lowest frequency is named the first formant (F1), the next the second formant (F1), the next the third formant (F3), and so on. Both vowels and consonants generate enough energy to evoke a response from the resonances in the vocal tract, resulting in formants (Plack, 2013). However, formants are likely to be more visible in vowel sounds because they are voiced, which in turn brings about resonances (Fant, 1960; Steven, 1980). In contrast, consonants often involve the co-ordination of voicing, aspiration (drawing in a breath), and frication (squeezing air through a small gap in the mouth), resulting in an anti-resonance effect (Stevens, 1980).

Formant patterns can be viewed using a wide band spectrogram and appear as dark, horizontal bands along the frequency scale (Plack, 2013). They are measured as amplitude peaks in the spectrogram and this gives an estimate of the vocal tract resonances (Reby & McComb, 2003). Compared to a narrow band spectrogram, these are created using a more coarse frequency analysis, the idea being to smear over a large enough band of frequencies to display the collection of frequency components that correspond to vocal tract formants (Plack, 2013). Figure 2.8 illustrates a wide band spectrogram for the spoken word *"heard"*.

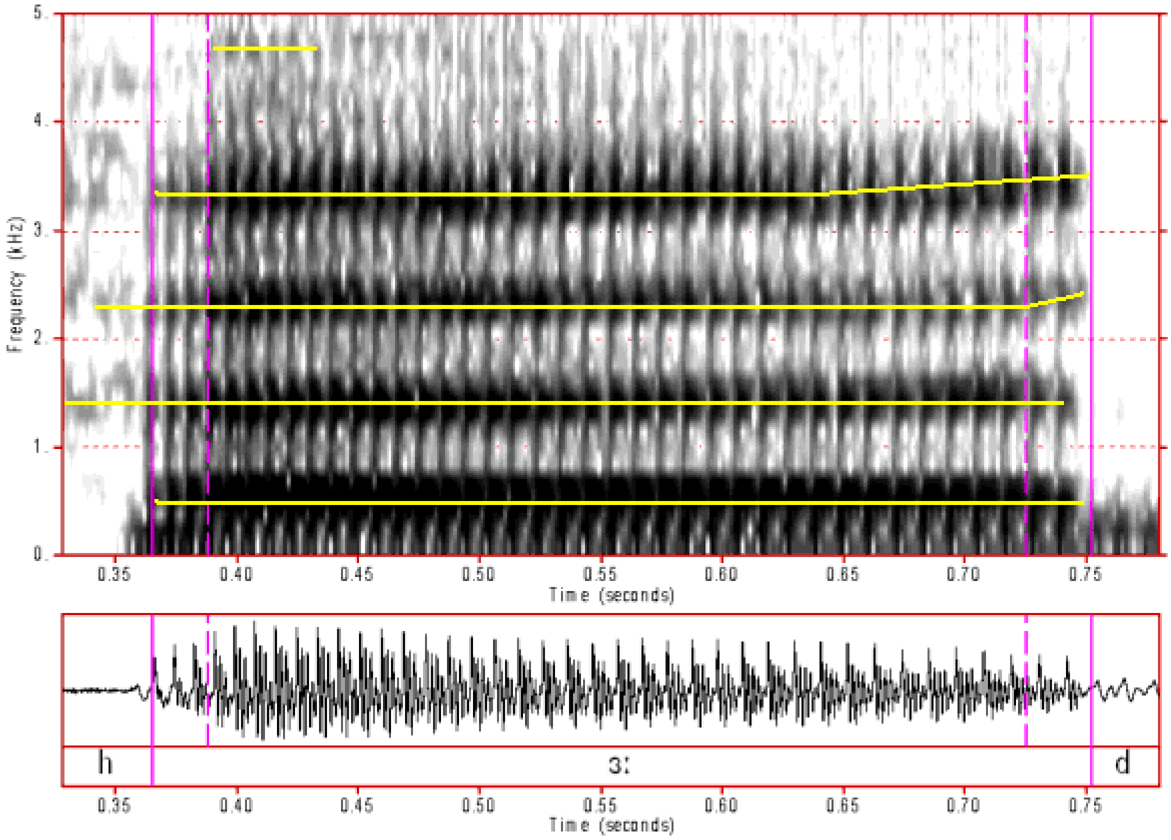


Figure 2.8: Wide band spectrogram of the vowel /ear/ (phonetic symbol /ɜ:/) in the word *"heard"*. Adapted from http://clas.mq.edu.au/speech/acoustics/ frequency/spectral/html. Dark black indicates greatest energy (amplitude) whereas white indicates least energy (amplitude). Formants are indicated by the yellow lines.

The first four formants (F1, F2, F3, and F4) can clearly be seen in Figure 2.9 and are highlighted using horizontal yellow lines. The first four formants are particulary apparent in the vowel sound /ear/, but F2 and F3 can also be seen in the consonant /h/. The wide bandwidth allows for excellent time resolution and is therefore able to capture rapid changes in amplitude that occur when the vocal folds vibrate (Plack, 2013). These can be seen as evenly spaced vertical lines in the voiced segments of the spectrogram and corrrespond to the individual frequency periods of the sound wave (Plack, 2013). Again, the sound wave and indvidual speech sounds for the utterance is also present, with the vertical pink lines dividing both the spectrogram and the sound wave into moments in time.

Each speech sound has a unique pattern (combination of formants) which allows listeners to classify and distinguish different sounds (Benavides et al., 2016). These patterns occur consistently no matter who the speaker happens to be. However, the frequencies at which they occur can differ as they are dependent on the F0 of the speaker (Davenport & Hannahs, 2010). Nevertheless, whilst the frequencies of the formants can change, they cannot change independently of other formants in that sequence, rather they appear in certain frequency combinations (Wells, 1962). At any one time, there may be a number of formants visible for a particular speech sound, however, the first three formants (F1, F2, and F3) are of most importance in determining which sound is heard (Wells, 1962). Collectively they are referred to as the formant pattern (Diehl, 2008). Formants above F4 are usually weak if visible at all, on a spectrogram, and often do not reveal any further information about the speech sound (Davenport, Davenport & Hannahs, 2010). Table 2.1 provides several examples of the typical formant pattern for a set of vowels.

Table 2.1 illustrates how the frequencies of the formants differ between adult males, adult females, and children. Female speakers typically display higher formant frequencies than males, and children display higher frequency formants than females, due to differences in vocal tract length (this is explained further in Chapter 3, Section 3.1.2.1.1). As illustrated in Figure 2.9, vocal tracts are longer in adult males than they are in females (Samuelsson, 2006). Vocal tracts are also longer in females than they are in children (Samuelsson, 2006).

Table 2.1: *Examples of the typical formant pattern for males, female, and children, for several different American English vowels. Adapted from http://feilding.net/sfuad/musi3012-01/html/lectures/017_instruments_IV.htm.*

	Males			**Females**			**Children**		
Vowel in…	**F1**	**F2**	**F3**	**F1**	**F2**	**F3**	**F1**	**F2**	**F3**
"beat"	270	2300	3000	300	2800	3300	370	3200	3700
"bit"	400	2000	2550	430	2500	3100	530	2750	3600
"bet"	530	1850	2500	600	2350	3000	700	2600	3550
"bat"	660	1700	2400	860	2050	2850	1000	2300	3300
"part"	730	1100	2450	850	1200	2800	1030	1350	3200
"pot"	570	850	2400	590	900	2700	680	1050	3200
"boot"	440	1000	2250	470	1150	2700	560	1400	3300
"book"	300	850	2250	370	950	2650	430	1150	3250
"but"	640	1200	2400	760	1400	2800	850	1600	3350
"pert"	490	1350	1700	500	1650	1950	560	1650	2150

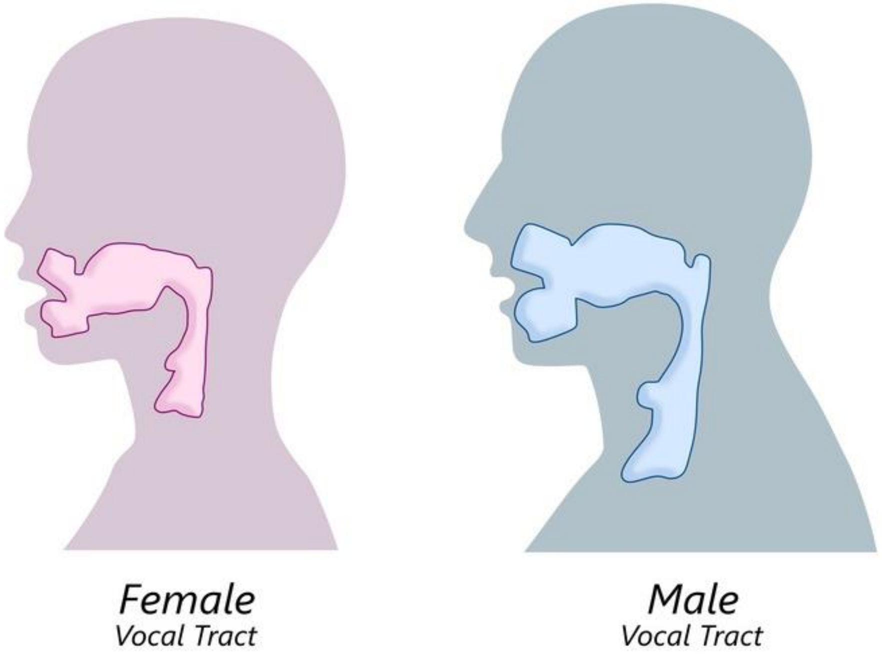


Figure 2.9: An illustration of the difference in the length of the vocal tract in males and females.

2.2 Properties of Speech

2.2.1 Linguistic and Paralinguistic Properties of Speech

As noted in Section 2.1, speech as the medium of human communication conveys information between a speaker and a listener on several layers (Laver, 1991). The information that is contained in the speech signal can be classified into two broad categories; linguistic and paralinguistic (or indexical) information (Levi & Pisoni, 2007; Rose, 2003).

Linguistic information refers to *what* is being said (i.e., the content of the utterance). The linguistic layer carries information about the symbolic content of the speaker's intended message (Laver, 1991). This includes phonological (i.e., units of sound), morphological (i.e., units of sound which form words), syntactic (i.e., combining words into sentences), and semantic (i.e., meaning of an utterance) information (Laver, 1991). Linguistic content in the speech signal serves a communicative purpose in that it conveys the message intended by the sender to make the receiver aware of something (Laver, 1979; 1989; 1994).

The voice also carries a vast amount of valuable biological and social information about the speaker (Belin, Fecteau & Bedard, 2004; Belin et al, 2011; Belizaire et al, 2007; Watson, Latinus, Noguchi, Garrod, Crabbe & Belin, 2014). This paralinguistic layer of communication is non-linguistic and non-verbal, and provides the listener with information about *who* is speaking. Thus, paralinguistic information refers to *how* we say something, rather than what is being said. Examples include information about the speakers physical identity (e.g., who is speaking, how old or young they sound, and whether they are male or female), and the affective state of the speaker (Honorof & Whalen, 2010; Imaizumi et al., 1997; Junger, Pauly, Brohr, Birkholz, Neuschaefr-Rube, Kohler, Schneider, Derntl & Habel, 2013). It has been suggested that the ability to extract such characteristic information from the voice constitutes a more primitive and universal non-linguistic mode of communication, and that typical listeners possess sophisticated cognition abilities for extracting and processing this speaker-related information (Belin, Zatorre, & Ahad, 2002; Latinus & Belin, 2012; Nakamura, 2001; Yovel & Belin, 2013). However, it should be noted that the paralinguistic markers are, in part, culturally determined and their respective interpretation must be learned (Laver, 1994).

Indexical properties of voice can be divided further into extrinsic and intrinsic properties (Laver, 1994). Extrinsic properties include those that are not under the speaker's control, yet can cause a change in the natural speech signal after it has been generated (Laver, 1994). Such properties include additive and background noise, different levels of noise in the external auditory environment, convolutive noise, competing talkers, and reverberation (e.g., echoes). Intrinsic properties however include those that are under the speaker's control, and refers to a collection of properties elicited in the generation phase of the voice (Laver, 1994; Levi & Pisoni, 2007). Intrinsic properties can include those that are less likely to vary and those that are subject to continuous change. Intrinsic properties that are less likely to vary include the speaker's dialect, accent, age, and sex. They also include physiological parameters, such as F0 (perceived as pitch). Intrinsic properties subject to continuous change include speaking effort (i.e., whether the speaker is talking loudly or soft), affective state (i.e., whether the speaker is happy, sad, angry etc.), speaking style (i.e., whether the utterance was a question or a statement), and speech rate (i.e., how fast or slow the speaker is talking) (Laver, 1994; Levi & Pisoni, 2007).

Both linguistic and paralinguistic information are carried simultaneously in the speech signal (Levi & Pisoni, 2007). In order for linguistic information to be conveyed, it has to pass through the speaker themselves. Attached to this linguistic information is indexical information about the speaker. Indexical information can therefore be thought of as the medium through which the linguistic message is conveyed (Levi & Pisoni, 2007). Figure 2.10 provides an example of the integration between linguistic and indexical properties in the acoustic waveform.

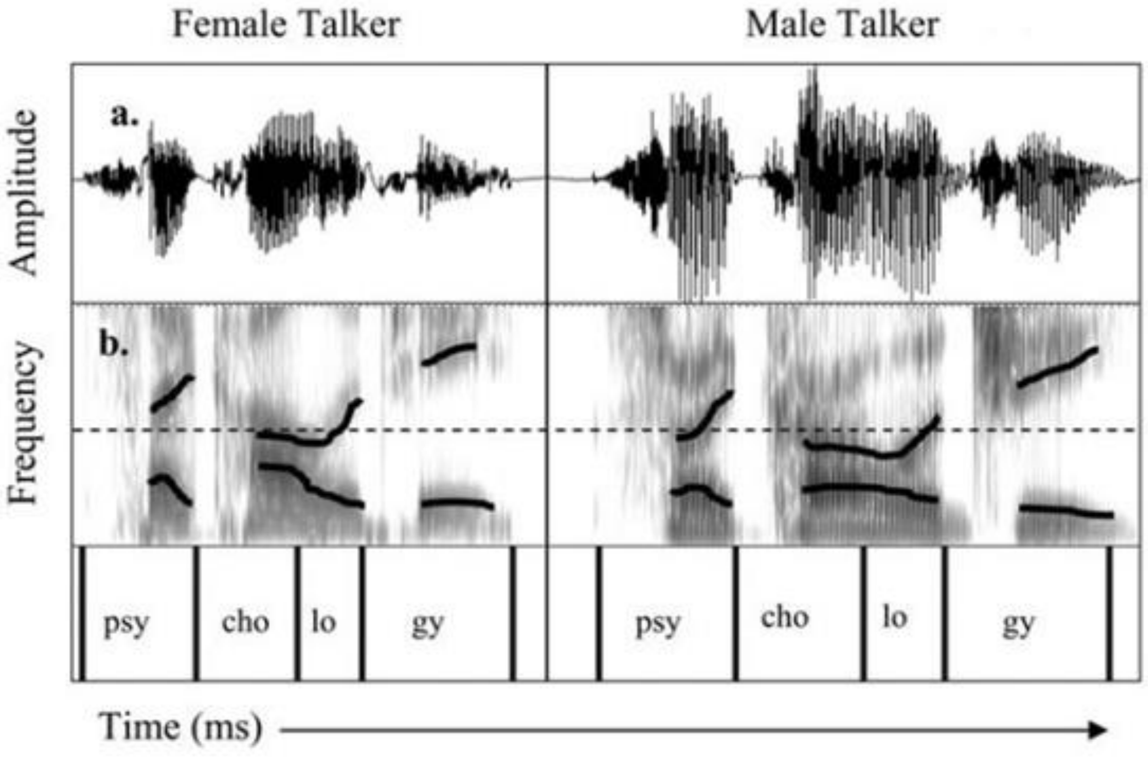


Figure 2.10: Sound wave (a) and spectrogram (b) of the spoken word *"psychology".* Adapted from Levi & Pisoni (2007). Dark lines in the spectrogram represent the first formant (lower curves) and the second formant (upper curves).

Figure 2.10 illustrates that the formant frequencies for the female talker are higher than those for the male talker. This provides indexical information about the speaker’s sex and is a result of the differences in vocal tract length. In contrast, the overall movement and relative locations of the formants provide the listener with linguistic information, and indicate that the speakers are saying the same utterance (Winters, Levi & Pisoni, 2011).

2.2.2 Acoustic Properties of Speech

Of the intrinsic paralinguistic properties listed above, some of these can also be considered acoustic properties of voice. Acoustic properties still provide characteristic information about the speaker, however they are also directly measurable from the speech signal (Choi, Hasegawa-Johnson & Cole, 2005). The speech signal contains two main features; temporal features and spectral features. Temporal features are time domain features of the speech signal and measurements include length of pauses (interval of speech where voicing is not present), minimum and maximum amplitude, and speech rate (Choi et al., 2005; Levi & Pisoni, 2007). Spectral features are frequency based features of the speech signal, and measurements include F0, rising and falling frequency, spectral flux (how quickly the spectrum is changing), spectral centroid (centre of mass of the spectrum), and spectral density (how the strength of a signal is distributed across different frequencies) (Levi & Pisoni, 2007).

2.2.3 Paralinguistic Properties and Their Acoustic Correlates

Human speech contains many acoustic cues that are indicative of a particular characteristic (Laver, 1991; Laver & Trudgill, 1991). Examples include the speaker’s age, sex, emotional state, and even their identity. The relationship between these paralinguistic properties and acoustic cues is usually a correlational one, where an increase or decrease in one leads to an increase or decrease in the other (Laver, 1991).

Acoustic properties of speech can correlate with more than one type of information (Belin, Fecteau, & Bedard, 2004). F0 is a spectral acoustic cue which can signal several physiological characteristics of the speaker, including the speaker's sex, age, height and weight, and also psycho/socio-logical characteristics, such as the speaker's affective state (Belin et al., 2004). In terms of speaker sex, observations across both the male and female F0 range emphasise a correlational relationship with F0. Specifically, a higher F0 is more likely to signal a female speaker, whereas a lower F0 is more likely to signal a male speaker (e.g., Assman et al., 2006; Coleman, 1976; Gelfer & Mikos, 2005). This topic will be addressed further in Chapter 3, Section 3.1.2.1.2.

In terms of speaker age, observations across both the male and female F0 range also suggest a correlational relationship with F0. For females, F0 has been found to decrease as the age of the speaker increases, whereas for males, F0 has been found to decrease until the speaker reaches approximately 40 to 50 years of age, at which point F0 gradually begins to increase until it reaches its peak at approximately 80 years of age (e.g., Benjamin, 1981; Chatterjee, Halder, Bari, Kumar, & Roychoudhury, 2011; Ferrand, 2001; Hollien & Shipp, 1972; Linville, 1996). Other examples include the speaker's affective state (F0 changes depending on the type of emotion elicited by the speaker) (e.g., Murry & Arnott, 1993; Scherer, 2003), the speaker's weight (e.g., Evans, Neave & Wakelin, 2006), and the speaker's height (e.g., Garddol & Swan, 1983; Puts, Apicella & Cardenas, 2012). Note that F0 has also been found to be inversely correlated with both weight and height (Hughes, Dispenza, & Gallup, 2004; Neave, 2006). This topic will be addressed further in Chapter 3, Section 3.1.3.3.

Speech rate is a temporal acoustic cue which can also signal physiological characteristics of the speaker, including the speaker's age, and psycho/socio-logical characteristics, such as the affective state of the speaker and their regional accent (Levi & Pisoni, 2007). In terms of speaker age, a faster speech rate is often perceived as being more characteristic of a younger speaker, whereas a slower speech rate is more characteristic of an older speaker (Shipp, Qi, Huntley, & Hollien, 1992). Observations across the age range of speakers emphasises a correlational relationship with speech rate, where increasing the speech rate also leads to an estimated increase in age (e.g., Ptack & Sander, 1966; Shipp et al., 1992). This topic will be addressed further in Chapter 3, Section 3.1.3.3.

Figure 2.11 provides an illustrative summary of the properties of speech that have been discussed so far.

2.3 Acoustic Variations in Speech

2.3.1 Inter- and Intra-Speaker Variations in the Voice

As noted in Section 2.2.2, acoustic properties of speech are variable. Variations in the acoustic properties of the voice exist between different speakers (also known as between speaker, or inter-speaker variation). Different speakers have different sounding voices because of physiological differences in the structure of speech mechanisms and the use of the vocal tract (Atkinson, 1998). Variations can also occur within the same speaker (also known as within-speaker, or intra-speaker variation) (Atkinson, 1998). Speakers rarely pronounce given words or phrases in an identical way on different occasions, even if the second utterance is produced in close succession (Hollien, 1990). The same speaker can sound different from time-to-time because of factors such as time of day, fatigue, intoxication (from alcohol or drugs), thought distractions, situational demands, changes in health and physical status, stress, a speaker's mood state, and a speaker's emotional state (Nolan, 2005; Saslove & Yarmey, 1980). These are all examples of *unintentional* modifications made to the voice. However, speakers can also choose to *intentionally* modify their own voice by means of disguise. Voice disguise refers to any intentional alteration, distortion, or deviation from the speaker's normal voice (i.e., the voice most typically produced by a speaker) (Rodman, 1998).

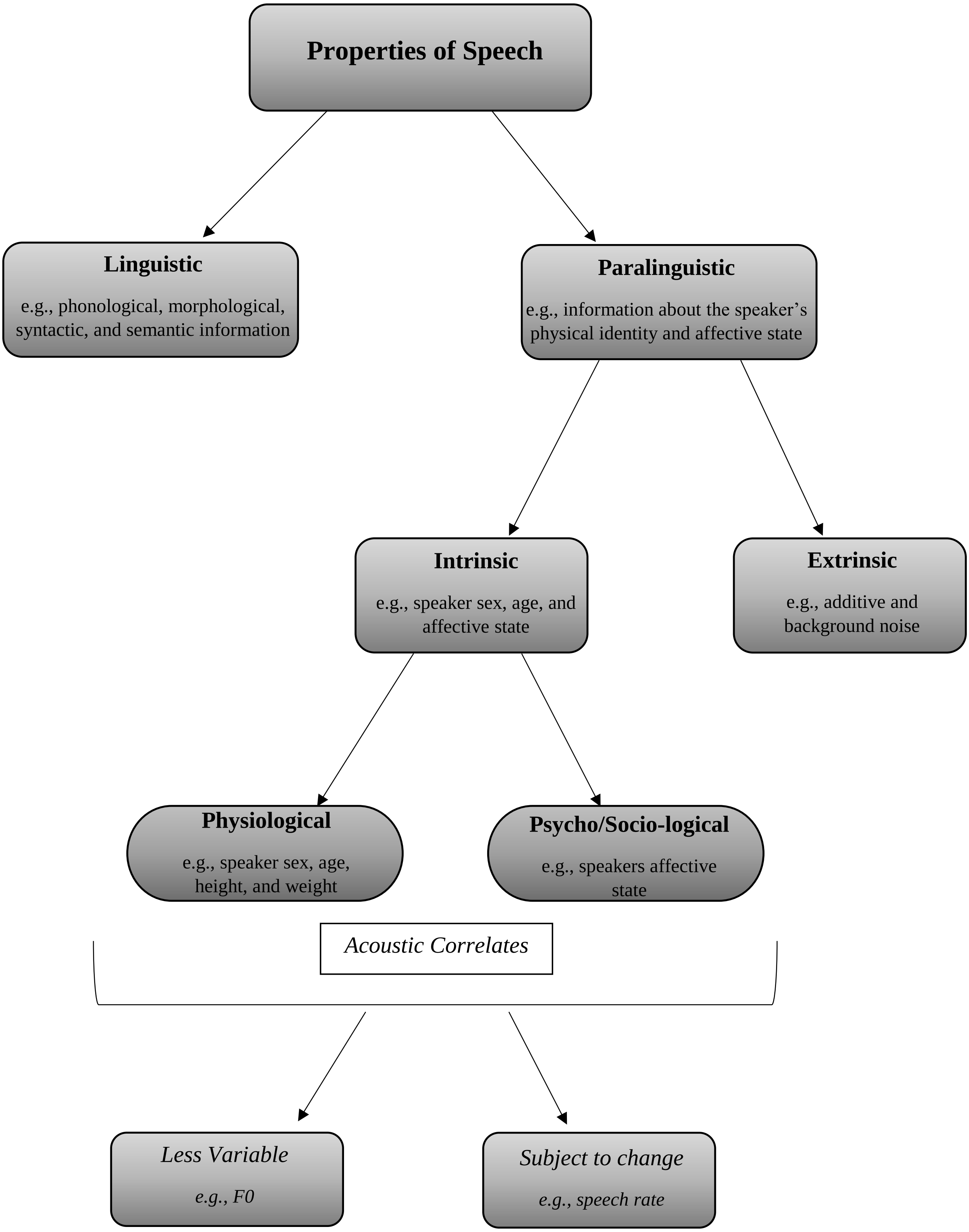

Properties of Speech
Linguistic
e.g., phonological, morphological, syntactic, and semantic information
Paralinguistic
e.g., information about the speaker's physical identity and affective state
Intrinsic
e.g., speaker sex, age, and affective state
Extrinsic
e.g., additive and background noise
Physiological
e.g., speaker sex, age, height, and weight
Psycho/Socio-logical
e.g., speakers affective state
Acoustic Correlates
Less Variable
e.g., F0
Subject to change
e.g., speech rate

Figure 2.11: A visual representation of the properties of speech, split into linguistic and indexical features, intrinsic and extrinsic indexical features, and acoustic features. Examples of these are also provided.

2.3.2 Controlling the Sound

Section 2.3.1 explained that the speech stream is a highly variable signal and that within-speaker variation exists. The same speaker can sound different from time-to-time because of both unintentional and intentional variations in the voice, and modifications in both F0 or speech rate will often occur as a result of these variations (Nolan, 2005). The following section will outline how a speaker can manipulate F0 or speech rate of their voice.

2.3.2.1 Control of Fundamental Frequency (F0)

The control of F0 is a complex interplay between respiratory control and the muscles in the larynx affecting vocal fold posture (Chhetri, Neubauer & Berry, 2012). The *body-cover* model is the predominant framework for understanding the control of F0 during speech (Hirano, 1974). It proposes that F0 is controlled by a change in the length (or strain) of the vocal folds and a change in the stress (or tension) of the tissues of the vocal folds (Hirano, 1974). The vocal folds are divided into two tissue layers with different mechanical properties (Story & Titze, 1995). The body layer consists of the thyroarytenoid (TA) muscle and deep collagen fibers, and the ‘cover’ layer consists of non-contractible tissues including the superficial and intermediate lamina propria layer, and the vocal fold epithelium (Hirano 1974). In this model, cover layer stiffness is primarily responsible for F0 control, and the TA and the cricothyroid (CT) muscles change the stiffness of the cover layer by altering its length (Chherti et al., 2012). Contraction of the CT muscles elongates and stiffens the cover layer, thus *increasing* F0, while activation of the TA muscles shortens the body layer while concurrently creating a slack in the cover layer, thus *decreasing* F0 (Chherti et al., 2012). Figure 2.12 provides an illustration of the main muscles in the larynx affecting vocal fold posture.

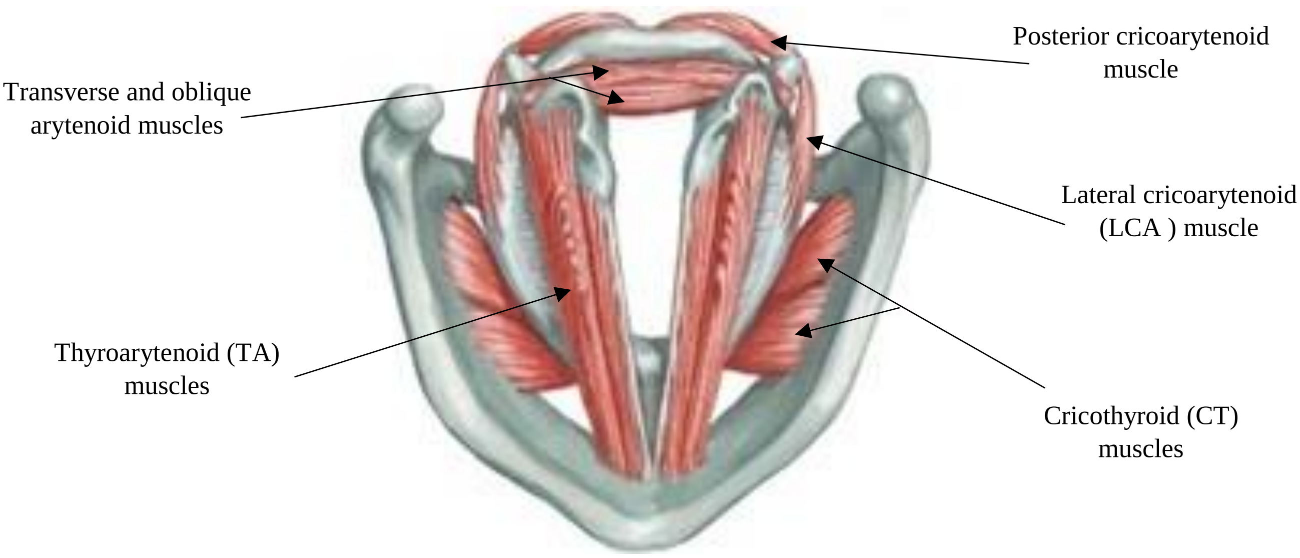

Posterior cricoarytenoid muscle
Transverse and oblique arytenoid muscles
Lateral cricoarytenoid (LCA) muscle
Thyroarytenoid (TA) muscles
Cricothyroid (CT) muscles

Figure 2.12: The main muscles in the larynx affecting vocal fold posture and controlling fundamental frequency (F0).

2.3.2.2 Control of Speech Rate

Three main explanations have been proposed for the changes in segmental timing that occur when speakers alter their rate of speech. Note, that a segment is any discrete unit (such as a consonant or vowel) that can be identified either physically or acoustically in the speech stream (Crystal, 2003). These explanations include the speed (i.e., the total rate of change of a movement trajectory) of selected articulatory movements, the distance over which the articulator moves during one or more speech gestures, and the relative timing (or phasing) of articulatory movements (Crystal, 2003). Each of these explanations will be briefly summarised below.

2.3.2.2.1 Rate Induced Variation in the Speed of Articulatory Movements

The first explanation as to how speakers alter their rate of speech suggests that speakers move the articulators an equivalent distance, but vary the speed of articulatory movements. Therefore, to *increase* speech rate, a speaker would move the articulators an equivalent distance, but *increase* the rate of the articulatory movements (Crystal, 2003). In contrast, to *decrease* speech rate, a speaker would move the articulators an equivalent distance, but *decrease* the rate of the articulatory movements (Crystal, 2003). Indeed, several studies of speech rate effects on articulatory movement speed have reported increased peak velocities of articulators with increased speech rates (Abbs, 1973; Adams, Weismer & Kent, 1993; Flege, 1988; Gay & Hiorse, 1973; Ostry & Munhall, 1985; Shaiman, 2001; 2002). This is consistent with the notion that to speak faster, a speaker must also move their articulators at a faster rate. However, others have indicated little or no evidence of changes in articulator velocities as a function of speaking rate (Bengueral & Cohen, 1974; Kent & Moll, 1972), and individual differences with regard to the occurrence of velocity changes have also been identified (Flege, 1988; Kuehn & Moll, 1976; Ostry & Munhall, 1985).

2.3.2.2.2 Rate Induced Variation in the Distance of Articulatory Movements

As an alternative to the above explanation, others have suggested that in order to control speech rate, speakers will maintain a similar speed of the articulatory movements, but vary the size of the articulatory movements so that more or less distance has to be covered (Crystal, 2003). Therefore, to *increase* speech rate, a speaker would *reduce* the size of the articulatory movements so less distance is covered, but maintain a similar speed of the articulatory movements. In contrast, to *decrease* speech rate, a speaker would *exaggerate (i.e., increase)* the size of the articulatory movements so more distance is covered, but maintain a similar speed of the articulatory movements (Crystal, 2003). Several studies have found a reduction in the distance of articulatory movements for faster speech rates (e.g., Kuehn & Moll, 1976; Lapointe, 2005). Furthemore, Gooze, Lapointe, and Murdoch (2003) found that eight out of ten participants reduced articulatory movement distances when speech rate was increased. Nevertheless, studies have also found speakers to demonstrate patterns of velocity increase and decrease in articulatory movements *in addition to* distance changes in articulatory movements when speech rate is increased (e.g., Abbs, 1973; Hertrich & Ackermann, 2000; Kent & Moll, 1972; Kuehn & Moll, 1976; Ostry & Munhall, 1985; Shaiman, 2001). This highlights the possibility of interactions between strategies used to change the rate of speech.

2.3.2.2.3 Rate Induced Variation in the Phasing of Articulatory Movements

The third explanation suggests that rate of speech can also be varied by changing the relative timing of successive articulatory movements, so that the overall duration of the event series is shortened (Crystal, 2003). This process is commonly referred to as *coarticulation*, or *coproduction* (Crystal, 2003). For example, when producing the first syllable in the word *"object"*, the speaker will have to lower the mandible for the vowel and also begin to approximate the lips to produce the bilabial plosive that follows (Berry, 2011). Note that a bilabial plosive is a consonant that is produced by stopping the airflow using the lips, teeth, or palate, followed by a sudden release of air (Berry, 2011). When the speaker begins the lip approximation earlier (relative to the jaw lowering), the overall duration of the sequence can be shortened, thus increasing speech rate (Berry, 2011). Measures that reflect the inter-articulator temporal overlap, or phasing, have been studied. Nevertheless, the literature is somewhat mixed, with reports that increasing speech rate results in increased overlap, no change in overlap, or decreased overlap (e.g., Abbs, 1973; Boyce, Krakow, Bell-Berti, & Gelfer, 1990; Byrd & Tan, 1996; Engstrand, 1988; Shaiman, 2001, 2002; Shaiman, Adams, & Kimelman, 1995).

2.4 Summary Conclusions

- ❖ Speech is a complex sound wave and is composed of several frequencies. The frequency components of complex sound waves are called harmonics.
- ❖ The first component in a harmonic series is the fundamental frequency (F0).

- Fundamental frequency (F0) refers to the lowest frequency, and therefore the lowest repeating component, of the complex sound wave. All remaining harmonics are integer multiples of the fundamental frequency (F0).
- Formants represent speech sounds that emphasise certain frequencies at higher amplitudes. Formants are more likely to be more visible in vowel sounds because they are voiced, which in turn brings about resonances in the vocal tract.
- Information that is contained in the speech signal can be classified into either linguistic or paralinguistic (indexical) information.
- Linguistic information refers to *what* is being said, whereas paralinguistic information in non-verbal and refers to *how* we say something. Paralinguistic information provides the listener with information about who is speaking.
- Acoustic properties of the voice provide paralinguistic information about the speaker and are directly measurable from the speech signal.
- Fundamental frequency (F0) is a spectral acoustic cue which can signal several physiological and socio-logical characteristics of the speaker, including sex, age, height weight, and the speakers affective state.
- Speech rate is a temporal cue which can also signal several physiological and socio-logical characteristics of the speaker, including age, regional accent, and the speakers affective state.
- Different speakers have different sounding voices because of the physiological differences in the structure of the speech mechanism and the use of the vocal tract.
- Variations in acoustic cues of the voice can also occur within the same speaker.
- The body-cover model is the predominant framework for understanding control of fundamental frequency (F0) during speech.

- Several explanations have been proposed for the control of speech rate, including the speed of selected articulatory movements, the distance over which the articulators move during one or more speech gestures, and the relative timing of articulatory movements.

CHAPTER 3. LITERATURE REVIEW:

SPEAKER PERCEPTION AND VOICE RECOGNITION MEMORY

3. Overview of Review

This chapter places the thesis within the wider context of existing literature. It begins by discussing differences that exist in fundamental frequency (F0) and speech rate of speakers of different identities, male and female speakers, and speakers of different ages. This section also reviews evidence that has considered manipulations in F0 or speech rate and how they affect perceptual judgements about the identity, sex, and age of the speaker. However, it comes to light that, despite this, the overall picture is still somewhat unclear. The findings also highlight several methodological issues with the studies that currently exist on these issues, which in turn, provides a rationale for investigating this topic further. The chapter then moves on to discuss the impact of manipulations in F0 or speech rate on recognition performance for these cues. It begins by reviewing evidence that has considered the effect of placing stimuli into distinct categories, by demonstrating how memory has been found to often reflect typical representations of a stimulus rather than specific features of those learned items (i.e., the accentuation effect). It becomes apparent that very few researchers have considered categorisation or accentuation effects in relation to voices, and highlights several methodological issues with the studies that do currently exist on this issue. Finally, the chapter considers other factors that might contribute to performance on a memory task for voice F0 and speech rate. It discusses the time course of echoic memory, and reviews research that has found people to be more accurate in a task that uses shorter intervals between presentations of the stimuli. It also discusses how memory for the voice may be somewhat easier if the sentence spoken is the same throughout the duration of the task. In discussing the existing research in detail, several gaps in knowledge emerge and serve as a framework on which the research questions in this thesis have been formulated.

3.1 Speaker Perception

When we hear someone speak, we do more than just understand the message it contains; we also make judgements about characteristics of the speaker based on the voice alone (refer to Chapter 2, Section 2.2.1 for further detail). For example, we try to ascertain who is speaking, how old somebody is, or whether the person speaking is male or female (Simpson, 2009). Human speech contains many acoustic cues that are indicative of a particular characteristic (Laver, 1991; Laver & Trudgill, 1991). The relationship between these paralinguistic properties and acoustic cues is usually correlational, where an increase or decrease in one property or cue leads to an increase or decrease in the other (refer to Chapter 2, Section 2.2 3 for further detail). As our physical characteristics change, how we sound also changes, and physical differences between speakers (i.e., inter-individual variation) may be reflected in consistent differences in how they sound. However, variations can also occur within the same speaker (i.e., intra-speaker variation), and these variations can occur both unintentionally (e.g., due to changes in the time of day, fatigue, situational demands, health, stress etc.) and deliberately (i.e., by means of disguise) (refer to Chapter 2, Section 2.3.1 for further detail).

Studies of the extent to which listeners can judge a speaker's physical characteristics are common in the voice literature, in part because of the inherent interest on the topic (Kreiman & Sidtis, 2011). Research has considered the perceptual cues utilised for decisions on speaker sex (e.g., Bachorowski & Owren, 1999; Lass, Hughes, Bowyer, Waters, & Bourne, 1976; Skuk & Schweinberger, 2013; Smith & Patterson, 2005), age, (e.g., Hartman & Danhauer, 1976; Smith & Patterson, 2005), size (e.g., Smith & Patterson, 2005), emotion (e.g., Bachorowski, 1999), or personality (e.g., Brown, Strong, & Rencher, 1974). These studies have often identified that manipulations in acoustic cues of the voice, and particularly in fundamental frequency (F0) and speech rate, can affect perceptual judgements for some of the characteristic information about the speaker. However, despite this history in perceptual speaker identification, the overall picture remaining is unclear (Sell, Suied, Elhilali, & Shamma, 2015). Furthermore, there are several methodological issues that make it difficult to determine the relevance of the findings (these will be expanded on further throughout this Chapter).

Physical characteristics such as the speaker's identity, sex, and age, are all biologically important, and the manner in which such information is transmitted has applied implications. Understanding what cues are perceptually important, and which do not produce noticeable changes in a person's voice, could provide insight into social relations of many kinds (Kreiman & Sidtis, 2011). For example, following a crime, the police will often ask the victim or witnesses to provide characteristic information about the voice of a suspect (Waller & Eriksson, 2016). Such descriptions are made frequently by victims and witnesses of crime who have encountered perpetrators under poor visual conditions (Yarmey, 2003; 2004). Testimonies may be based on observations in the dark, when the perpetrator is masked or wears a disguise, when the victim is blindfolded, or where an offence is committed over the telephone (Sherrin, 2014). In such cases, descriptions may be based solely on acoustic information from the voice. It is important for law enforcers to have knowledge about the accuracy with which listeners judge a speaker's physical characteristics, and the grounds on which estimations about a speaker's physical characteristics are made as this could aid in profiling criminals where only voice information is available, and enhance the accuracy and relevance of testimony in court.

The following section will address the extent to which manipulations in F0 or speech rate can affect perceptual judgements about the characteristics of the speaker. The section will begin by considering the role of F0 and speech rate as cues to the perceived identity of the speaker. It will then move on to discuss perceptions of speaker sex. The physiological, anatomical, and behavioural differences in the F0 of male and female voices will be described before reviewing the research that has considered the role of F0 as a cue to perceived speaker sex. This section will also report the actual and perceived differences in the speech rate of male and female voices. Finally, the section will consider the role of F0 and speech rate in perceptions of speaker age. The structural, functional, hormonal, and behavioural differences in the F0 and speech rate of male and female voices will be addressed before discussing the research that has examined F0 and speech rate as cues to perceived speaker age.

3.1.1 Perceptions of Speaker Identity

3.1.1.1 Fundamental Frequency (F0) and Speech Rate as Cues to the Perceived Identity of the Speaker

Whilst numerous studies in the past have examined the ability of listeners to accurately identify human speakers (Kreiman & Sidtis, 2011), very few have attempted to determine the contribution of acoustic cues of the voice and how they affect perceptions of speaker identity. In recent years, there have been a few examples of research directed at examining the effect of acoustic cues through direct manipulation of the signals (Sell et al., 2015), and there is some psychoacoustic evidence to suggest that the identification of the speaker may rely on the extraction of such information (Belin, Fecateau, & Bedard, 2005). However, such work has primarily focused on the effect of these manipulations on judgements of speaker similarity (i.e., ‘how similar does this voice sound to the voice you previously heard?’) rather than speaker identity (i.e., asking listeners to determine whether two utterances with differing degrees of manipulation were produced by the same speaker). Thus, at present any conclusions about the role of acoustic cues on perceptions of speaker identity are rather limited.

Of the few studies that do exist, research suggests that F0 may be a particularly important cue when making perceptual judgements about the identity of the speaker. Indeed, Lavner, Gath, and Rosenhouse (2000) manipulated F0 recordings of the vowel sound /a/ spoken in isolation by eight different male speakers. For each speaker, a reference sound of the spoken vowel was set at a frequency of 100 Hz. Each reference sound was then manipulated at three different frequency's (120 Hz, 140 Hz, and 180 Hz). The reference sound was presented successively with one of the manipulations and listeners were asked to judge whether the sounds were spoken by the same speaker or by a different speaker. The measure of (dis)similarity was the percentage of time listeners judged the voices as being a different identity. The results showed that greater manipulations in F0 (i.e., 180 Hz) led listeners to perceiving the voices as being spoken by different speakers at a significantly greater rate than when smaller manipulations in F0 were made (i.e., 120 Hz). However, recognisability of each voice was influenced differently by manipulations in F0. This suggests that the feature set utilised by the listeners varied with the speaker. Nevertheless, given the study only used male speakers, it is difficult to determine whether manipulations in F0 would also contribute to perceptions in the identity of female speakers. Although there is no reason to suggest that any differences would exist between the cues used to make judgements about speaker identity for male and female speakers, this still needs to be tested empirically to indeed determine whether this is the case.

In another study, Kuwabara and Takagi (1991) manipulated upward and downward shifts in F0 in two speakers uttering nonsense words. The speakers of the manipulated utterances were identified by three listeners who were familiar with the original speakers. The results showed that correct identification (i.e., saying that two voices were the same speaker) was reduced to chance performance when F0 was changed by 4.5 semitones (approximately 30% change in overall mean F0). An advantage of this study is that the researchers used utterances rather than isolated vowel sounds (as in Lavner et al., 2000). It could be argued that isolated sounds are unlikely to be heard, or used, by listeners when attempting to identify the speaker. The results of the study are therefore more generalizable to a real-world environment. Nevertheless, the participants recruited in the study are unlikely to be representative of the target population because only three people were used. Furthermore, both Lavner et al. (2000) and Kuwabara and Takagi (1991) used speakers familiar to the listener. Whilst this distinction may seem nuanced, familiar and unfamiliar speaker identification have been shown to be measurably different tasks (Yarmey, Yarmey, & Parliament, 2001) that utilise different regions of the brain (van Lancker & Kreiman, 1987), and studies have not yet shown whether acoustic features utilised by a listener are similar or different for familiar or unfamiliar speakers.

Gaudrain et al., (2009) examined the relationship of F0 and speaker similarity by manipulating F0 of consonant-vowel syllables spoken in isolation by an unfamiliar speaker. Listeners were asked to rate how similar the manipulated versions were to a reference sound (i.e., the unmanipulated version). The measure of (dis)similarity was the percentage of time listeners judged the voices as being different. The results showed that listeners believed the utterances sounded different 50% of the time or more when voices were manipulated in F0 by 25%. An advantage of this study is that the researchers used an unfamiliar speaker rather than a familiar speaker. However, the stimuli set was small as only one speaker was used, making it difficult to determine whether the results would replicate across other voices. Additionally, judgements of speaker similarity were made, thus making it difficult to determine whether the same manipulations made in F0 could also change the identity of the speaker.

To determine the acoustic cues responsible for perceptual judgements in unfamiliar speaker identity, Sell et al. (2015) used unfamiliar speakers to establish whether a listener’s ability to discriminate between utterances consisting of the same spoken words was affected by manipulations in F0. Six male speakers were used for experimentation. Each of the six voices were manipulated by resynthesizing mean F0 to the overall mean F0 of the six voices (113.27 Hz). The unmodified versions were presented successively with the resynthesised versions, and listeners were asked to determine whether the utterances were spoken by the same speaker or a different speaker. The results showed that changes in F0 did affect the ability of listeners to correctly identify speakers. However, the listeners were still able to consistently perform to a high level, and always above chance. Susceptibility to whether changes in F0 affected speaker identity was also found to be partially dependent on the specific speaker; changes in F0 affected perceptions of speaker identity for some speakers more than for others. The researchers concluded that listeners use features beyond F0 in speaker identification and can perform speaker tasks without that information (Sell et al., 2015). The results of this study are likely to be more generalisable to the real world given that spoken sentences were used. However, the study only used male speakers in their stimuli set, making it difficult to determine whether the results are also applicable to female speakers. Furthermore, only one modification in F0 was made to the voices rather than a series of changes, making it impossible to establish at what point listeners perceive voices as sounding like a different speaker. Moreover, the utterances were resynthesized to the overall mean F0 of the six voices. It is possible that the manipulations were too small to determine whether F0 could change the perceived identity of the speaker.

The studies discussed so far have used controlled manipulations in F0. Mathur, Choudhary, and Vyas (2016) made use of speakers naturally varying F0 by asking them to disguise their voice. To do this, they asked the speakers to increase and decrease the frequency of their normal spoken voice. Values of F0 in disguise by lowering the frequency of the voice were found to be significantly different compared to their respective control samples (i.e., voices spoken normally), but only for male speakers. Values of F0 in disguise by increasing the frequency of the voice were found to be significantly different compared to their respective control samples for both male and female speakers. Such findings are insightful as they imply that attempts to disguise voices in the real world may successfully alter perceptions of speaker identity. However, the study did not determine whether such attempts to disguise the voice would result in listeners perceiving the speaker as being a different identity, thus, any conclusions drawn from this experiment so far are only speculative.

To the best of this authors knowledge, there is only one study that has explored manipulations in speech rate on perceptions of speaker similarity. Starting with a control voice, Brown (1981) varied both F0 or speech rate by increasing and decreasing them each by 20%. Listeners were asked to judge the similarity of the control voice with the manipulated versions. The results showed that manipulations in both F0 and speech rate affected similarity judgements. Nevertheless, the study only used one voice making it difficult to determine whether the results are replicable to other voice stimuli. Furthermore, the findings cannot establish whether manipulations in F0 and speech rate alter perceptions of speaker identity given that judgements of similarity were used rather than judgements in identity.

In summary, it appears that F0 is an important cue in determining the identity of the speaker. Manipulations in F0 appear to affect perceptions of speaker identity, and changes the likelihood that listeners will accurately determine the identity of the speaker. The evidence for speech rate affecting perceived identity of the speaker is more limited. However, the results so far tend to suggest that manipulations in this cue may also be of use when attempting to determine the identity of the speaker. Nevertheless, given that the evidence is somewhat limited, and that there are several methodological issues with the existing research, it is difficult for any substantive conclusions to be made. The subsequent experimentation carried out in Experiment 2 (Chapter 5) seeks to fill this gap in the literature.

3.1.2 Perceptions of Speaker Sex

3.1.2.1 Fundamental Frequency (F0) and Speaker Sex

3.1.2.1.1 Physiological and Anatomical Differences in the Fundamental Frequency (F0) of Male and Female Voices

Humans exhibit large sexual dimorphism in vocalisations and vocal anatomy (Puts, Apicella & Cardenas, 2011). These differences between male and female speech are known to be a decisive clue in the identification of sex from speech (Perry, Ohde, & Ashmeand, 2001). Mean F0 is the major cross-sex acoustic difference between adult female and male voices (Gelfer, & Mikos, 2005; Pepiot, 2014). Adult male speakers usually have a lower F0, and therefore a lower perceived pitch, than adult female speakers. The voiced speech of an adult male will have a F0 between 85 Hz to 180 Hz, whereas an adult female will have an F0 between 165 Hz to 255 Hz (Baken, 1987; Titze, 1994). Typically, the F0 for an adult male is around 120 Hz, while a typical F0 for an adult female is around 200 Hz (Perry, Ohde, & Ashmead, 2001). Studies suggest that F0 is one of the most decisive cues in the perception of speaker sex from the voice (Pepiot, 2015). For example, research has reported significantly better identification of speaker sex from phonated than from whispered vowels (Bennett & Montero-Diaz, 1982; Lass, Hughes, Bowyer, Waters, & Bourne, 1976). In whispering, the vocal folds in the larynx do not vibrate, but are held close together (Martin, 2015). Consequently, phonated sounds provide information about the F0 of the speaker, whereas whispered sounds do not. Furthermore, studies of the vocal characteristics of male-to-female transgendered individuals who are successfully perceived as female have consistently shown that increasing F0 is of primary importance when trying to shift the perception of the voice from male to female (Gelfer & Schofield, 2000; Spencer, 1988; Mount & Salmon, 1988; Wolfe, Ratusnik, & Smith, 1990).

Several studies have also brought to light other sex acoustic differences in the voice, particularly formant frequencies. For example, Schwartz and Rine (1968) and Bennett and Montero-Diaz (1982) reported near-perfect judgements of speaker sex from whispered vowels. Speaker sex can also be judged from isolated voiceless fricatives (Ingemann, 1968) and from sine wave replicas of short sentences (Fellowes, Remez, & Rubin, 1997). Sine wave replicas, being aperiodic, do not have F0 in the traditional sense. Therefore, perceptions of speaker sex for these utterances is almost certainly related to differences in formant frequencies (Hillenbrand & Clark, 2009). The frequency of formants will differ depending on the vowel being spoken, however, typically formant frequencies are about 20% higher in females than in males (Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952) (refer to Chapter 2, Section 2.1.3.5 for further detail).

Between-speaker variations in these acoustic parameters can mainly be accounted for by the anatomical and physiological differences in vocal fold size and vocal tract length that arise during puberty (Fant, 1966). Vocal fold size primarily determines F0 of voiced speech, whereas vocal tract length is what determines the frequency of vowel formants (refer to Chapter 2, Section 2.1.3.5 for a more detailed discussion). Prior to adolescence, no significant sex differences in vocal fold size or vocal tract length have been identified between males and females (Cruttenden, 1986; Fitch & Giedd, 1999; Hirano, Kurita, & Nakashima, 1983; Perry, Ohde, & Ashmead, 2001). However, this changes considerably at puberty and during the transition into adolescence and adulthood. The cause of this change in vocal fold size and vocal tract length in males and females are largely due to hormonal changes during puberty. Girls begin puberty at approximately 10-11 years of age, and complete puberty by 15-17 years of age, whereas boys begin puberty at approximately 11-12 years of age, and complete puberty by 16-17 years of age (Cavanaugh, 2011). The larynx is particularly responsive to the sex hormones which include androgens, such as testosterone and dihydrotestosterone (DHT), progesterone, and estrogen (Kadakia, 2013). At puberty, elevated testosterone levels acting through androgen receptors in the vocal folds causes them to grow longer and thicker in males, however, no comparable changes occur in females (Fitch & Giedd, 1999; Harries, Hawkins, Hacking & Hughes, 1998; Newman, Butler & Hammond, 2000; Puts, Hodges, Cardenas & Gaulin, 2007). Reasons for this are unclear, but several researchers have suggested that both inter- and intra-sexual selection shaped men’s voices (Collins, 2000; Feinberg, Jones, Little, Burt & Perrett, 2005; Puts, 2005). This increase in size and thickness of the vocal folds causes them to vibrate more slowly and at approximately half the F0 of females during phonation (Kadakia, Calson & Sataloff, 2013; Puts, Apicella & Cardenas, 2011). By adulthood, a distinct difference in the size of the vocal folds between adult males and females is apparent (Cruttenden, 1986; Hughes & Rhodes, 2010; Klatt & Klatt, 1990; Titze, 1989; Ohde, Sharf & Jacobson, 1992; Perry et al, 2001). Adult male vocal folds are typically 60% longer and between 1.75cm and 2.5cm in length, while female vocal folds are between 1.25cm and 1.75cm in length (Thurman & Welch, 2000; Titze, 1994). Adult F0 is attained at an average of 15 years of age in females, and 16 years of age in males (Aronson & Bless, 2011).

Under the influence of androgens, the larynx also grows and descends in the neck in both sexes. This is however more dramatic in males than it is in females, resulting in a longer vocal tract and lower, more closely spaced formant frequencies (Fant, 1970; Fitch Giedd, 1999). Whilst androgens are present in females, their effect is not as noticeable until after the menopause. Instead, females mature in response to an increase in progesterone and estrogen. This results in a smaller decrease in frequency, about a third of an octave (Anderson & Sataloff, 2004). The average length of an adult female vocal tract is about 14.5cms (Simpson, 2009). In adult males it is about 17 to 18cms (Simpson, 2009).

3.1.2.1.2 Behavioural Differences in the Fundamental Frequency (F0) of Male and Female Voices

Acoustic differences between male and female voices are also influenced by behavioural factors (Perry, Ohde, & Ashmead, 2001). Indeed, research has suggested that humans are capable of producing speech patterns appropriate to the sex we identify with (Simpson, 2009). For example, it has been noted that adult males will often speak with an unnaturally lower F0, and females will often speak at an unnaturally higher F0 in order to conform to stereotypical views of vocal production characteristics (Sachs, Lieberman, & Erickson, 1973). Furthermore, when listeners are asked to describe stereotypical male and female voices, they typically expect male voices to be deep, demanding and loud, and female voices to have good enunciation, high pitch, and a fast and variable speaking rate (Kramer, 1977). Such evidence suggests that listeners have an expected set of characteristic cues that they use to judge a speaker's sex from a voice sample.

3.1.2.1.3 The Gender Ambiguous Range

Although characteristics of vocal folds and vocal tracts are appreciably different between male and female speakers, they do overlap to some extent (Sokhi, Hunter, Wilkinson, & Woodruff, 2005). The achievable range of F0 in male and female speakers overlaps in the F0 range of approximately 135 to 181 Hz (Andrews & Schmidt, 1997; Gelfer & Schofield, 2000; Henton, 1995). This range has been referred to as the gender ambiguous range of frequencies. In this range, the decision on a male or female voice depends on other parameters, like visual information or prosodic characteristics (e.g., intonation, stress, and rhythm of speech sounds), in order to correctly determine the sex of the speaker (Gelfer & Schofield, 2000; Oates & Dacakis, 1997). This gender ambiguous range is centred around a gender-cut off F0 of approximately 160 Hz, with voices below 160 Hz being assigned to males (Oates & Dacakis, 1997); Spencer, 1988). Especially in this range, formant frequencies get a high importance as additional determining cues in identifying the sex of the speaker (Sokhi et al, 2005; Titze, 1989).

3.1.2.1.4 Fundamental Frequency (F0) as a Cue to the Perceived Sex of the Speaker

Given such sexual dimorphism in human speech characteristics, researchers have attempted to investigate the role of F0 in perceptual judgements of speaker sex. Early studies typically separated the contribution of F0 and formant frequencies by having male and female speakers produce vowels using an electrolarynx as the vibrating source instead of the larynx (e.g., Coleman, 1976). This device produces a buzzing sound at a constant frequency. When held against the neck it provides an acoustic source that excites the vocal tract in place of laryngeal vibrations. Experimentally, this alternative sound source makes it possible to combine the F0 of one sex with the formant frequencies of the other. Results using this technique have shown that when male appropriate F0 is paired with male appropriate formant frequencies, the sex of the speaker was correctly identified as male 98% of the time. However, when female appropriate F0 is paired with female appropriate formant frequencies, the sex of the speaker was correctly identified only 79% of the time. When male F0 was paired with female formant frequencies, the speaker was identified 67% of the time as male. Nevertheless, when a female F0 was paired with male formant frequencies, the speaker was still identified as male 70% of the time (Coleman, 1971). The findings suggest that F0 is a robust cue in determining the sex of the speaker and that it serves as a more compelling cue to speaker sex than formant frequencies.

A drawback of using an electrolarynx is that the resulting stimuli often sound monotone and highly unnatural (Kreiman & Sidtis, 2013). To overcome this problem, several studies have applied the use of synthesised speech to determine the relative contribution of F0 and formant frequencies to perceived speaker sex. For example, Whiteside (1998) used a matched/mismatched perceptual procedure similar to Coleman (1971) using synthesised vowels. When male-appropriate F0 was paired with male-appropriate formant frequencies, the speaker was identified as male 97.2% of the time. When female-appropriate F0 was paired with female-appropriate formant frequencies, the speaker was identified as female 85% of the time. When male-appropriate F0 was paired with female-appropriate formant frequencies, the speaker was identified as male 93.8% of the time. A female-appropriate F0 paired with male-appropriate formant frequencies resulted in the speaker being identified as female 74.6% of the time. A similar pattern of findings have also been found in other studies (e.g., Gelfer & Mikos, 2005). Such research suggests that F0 is more likely to be the dominant cue in the identification of the sex of the speaker compared with formant frequencies. The results also indicate that male cues seem to be more perceptually salient than female cues, and particularly in the Coleman (1976) study, there appeared to be a bias toward the perception of a speaker as male when any male characteristics was present.

The findings from the studies reviewed so far are however limited. Indeed, such studies made use of only isolated vowels, making it difficult to determine whether they would generalise to whole words or complete sentences. More recent work has focused on the use of spoken sentences in determining the relative contribution of F0 and formant frequencies in determining speaker gender. For example, Assman, Nearey, and Dembling (2006) increased and decreased the F0 and formant frequencies of sentences spoken by males and females. Listeners were asked to rate the signals on a continuous scale ranging from ‘clearly masculine’ to ‘clearly feminine’. The results showed that sentences with low F0 and formant frequencies were perceived as more masculine, while sentences with high F0 and formant frequencies were more feminine. However, ratings of masculinity for signals with downward frequency shifts were more pronounced than ratings of femininity for signals with equivalent upward shifts.

Sentences with mismatched F0 and formant frequencies were assigned ratings near the midpoint of the range, indicating gender ambiguity. The researchers also found that even with equivalent F0 and formant values, signals synthesised from sentences originally spoken by males were more likely to be heard as masculine than were signals originally spoken by females. Conversely, signals synthesised from sentences originally spoken by females were more likely to be heard as feminine.

The findings so far appear to suggest a perceptual advantage for male speech in tasks that involve perceptions of speaker sex. In an attempt to explain why this effect is so prevalent, Owren, Berkowitz, and Bachorowski (2007) argue the following:

> "because sexual selection leads males to diverge from the 'default' female form (i.e., physiological changes in speech structures that occur during puberty), adult male voices can be considered 'marked' by the sexually selected features of lowered F0 and formant frequencies. It therefore follows that listeners should hear talker sex somewhat more easily in male than in female voiced sounds. Specifically, the presence of critical features of 'maleness' (low F0, low formants) virtually guarantees that the talker is an adult male. However, their absence does not unequivocally imply that the talker is an adult female." (p. 930)

However, this male advantage is not always apparent in the literature. For example, Gelfer, and Bennett (2013) manipulated the F0 of sentences spoken by both males and females to F0's typical of average males, average females, and in an ambiguous range. The results indicated that female speakers were perceived as female even with an F0 in the typical male range. However, for male speakers, perceptions of speaker sex were less accurate at F0's of 165 Hz or higher (i.e., the lower bound cut off for the typical female F0 range for voiced speech). These findings appear to suggest the opposite; that perceptions of speaker sex are *more* accurate for female voices than they are for male voices.

In other work using the matched/mismatched perceptual procedure, Hillenbrand and Clark (2009) found that talker sex was conveyed almost perfectly (99.6%) for both male and female voices when the voices were unmodified. However, perceived talker sex shifted rather strongly from male to female when F0 and formant frequencies were shifted up (81.9% of the time), and to a nearly equal degree from female to male when both F0 and formant frequencies were shifted down (82.1% of the time). On a substantial majority of trials, shifts in only F0 or formant frequencies were ineffective in changing perceived speaker sex. By itself, F0 was more effective in shifting perceived talker sex than shifting formant frequencies, although increasing the F0 of male voices to a female-appropriate F0 was more effective at changing the perceived sex of the speaker to female (34.3% of the time) than decreasing the F0 of female voices to male-appropriate F0 (19.1% of the time). Increasing the formant frequencies of male voices to female-appropriate formant frequencies was slightly more effective at changing perceived speaker sex to female (18.9% of the time) than decreasing the formant frequencies of female voices to male-appropriate formant frequencies (11.7% of the time), although in the main, shifting perceptions of identity were small using formant frequencies. It should be noted that these effects were smaller for sentence stimuli than they were for vowels, pointing to the importance of articulatory and prosodic cues in the perception of speaker sex. Indeed, in approximately 18% of the trials using spoken sentences, utterances retained their original perceived speaker sex despite substantial shifts in both F0 and formant frequencies. Therefore, whilst it is evident that F0 and formant frequencies are important to perceptions of speaker sex, other cues are also used. Furthermore, the results of both the Gelfer, and Bennett (2013) and Hillenbrand and Clark (2009) studies demonstrate that cues more typical of female speakers are just as likely to change perceptions of speaker sex, if not more so, as cues more typical of male speakers.

In summary, the results suggest that manipulations in both F0 and formant frequencies are more effective at changing the perceived sex of the speaker than manipulations of either F0 or formant frequencies alone. For isolated vowels, F0 is often the most important cue to speaker sex, however, formant frequencies become increasingly important when sentence stimuli are used. Correctly recognising the sex of the speaker is usually easier when spoken sentences are used compared to the use of isolated vowels, suggesting that other cues are important in determining the sex of the speaker. In the main, there does appear to be some evidence to suggest a male-advantage in the perception of speaker sex. However, the evidence for this is not always clear cut, and in some cases, a female-advantage has also been found.

3.1.2.2 Speech Rate and Speaker Sex

3.1.2.2.1 Stereotypical Opinions About the Speech Rate of Male and Female Voices

Potential male-female differences in speech rate have also been investigated, although somewhat less extensively than they have for F0. Research has shown that people believe females speak at a faster rate than males (Weirich & Simpson, 2014). This belief is quite pervasive and has been given credence in the scientific literature, with some even making the unsubstantiated claim that females speak on average faster than males (e.g., Brizendine, 2006). Several researchers have investigated the acoustic parameters that correlate with speech rate to determine what might account for the persistence of this stereotype. For example, Bond and Feldstein (1982) investigated the effect of frequency on speech rate. Listeners were asked to rate the perceived frequency and speech rate of electronically altered sets of spontaneous speech with a duration of 20 seconds varying in frequency. Their results showed that a faster speech rate was perceived with an increase in F0. Given that the F0 of a typical adult female is higher than that of an adult male, this could be one contributing factor to the belief that females speak faster than males.

Others have identified the importance of dynamic F0 and sentence duration in the perception of speech rate. Dynamic F0 is the variable increase and decrease in frequency during a spoken utterance. Listeners generally perceive utterances spoken with a more dynamic F0 as sounding more complex and longer than those spoken with a flat F0 (Cumming, 2011; Fougeron, Kuehnert, Imperio & Vallee, 2010; Lehiste, 1976). Longer sentence durations are a significant predictor of speech rate so that longer phrases, containing more syllables compared to shorter ones, are generally spoken at a faster rate owing to anticipatory shortening of the syllables (Quene, 2008). Anticipatory shortening is when the duration of the vowel in the first syllable of a word is shortened when the number of syllables that follow the stressed syllable increases. Not only have female speakers been found to vary F0 of a spoken utterance more frequently than male speakers, they have also been found to have longer sentence durations compared to male speakers (e.g., Weirich & Simpson, 2014). Thus, it is likely that listeners will perceive the rate of spoken utterances of a female speaker as faster than the utterances of a male speaker.

Perceived speech rate may also be influenced by acoustic vowel space size. Acoustic vowel space is a co-ordinate representation of the locations of an individual’s vowels, according to two key properties of the speech signal, frequency formants F1 and F2. As previously noted, the cavities and moving articulators of the vocal tract act as spectral filters during speech, which is what gives rise to the different characteristic formant structures that can be heard in vowel space (refer to Chapter 2, Section 2.1.2.2). Acoustic vowel space can be mapped out by taking measurements from people reading key vowels of their language. These spaces vary from person to person due to sociocultural and biological factors, however, cross-linguistically it has been demonstrated that the average female acoustic vowel space tends to be larger than the average male acoustic vowel space (e.g., Diehl, Lindblom, Hoemeke, & Fahey, 1996; Hillenbrand, Getty, Clark, & Wheeler, 1995; Whiteside, 2001; Simpson, & Ericsdotter, 2007). This is illustrated in Figure 3.1.

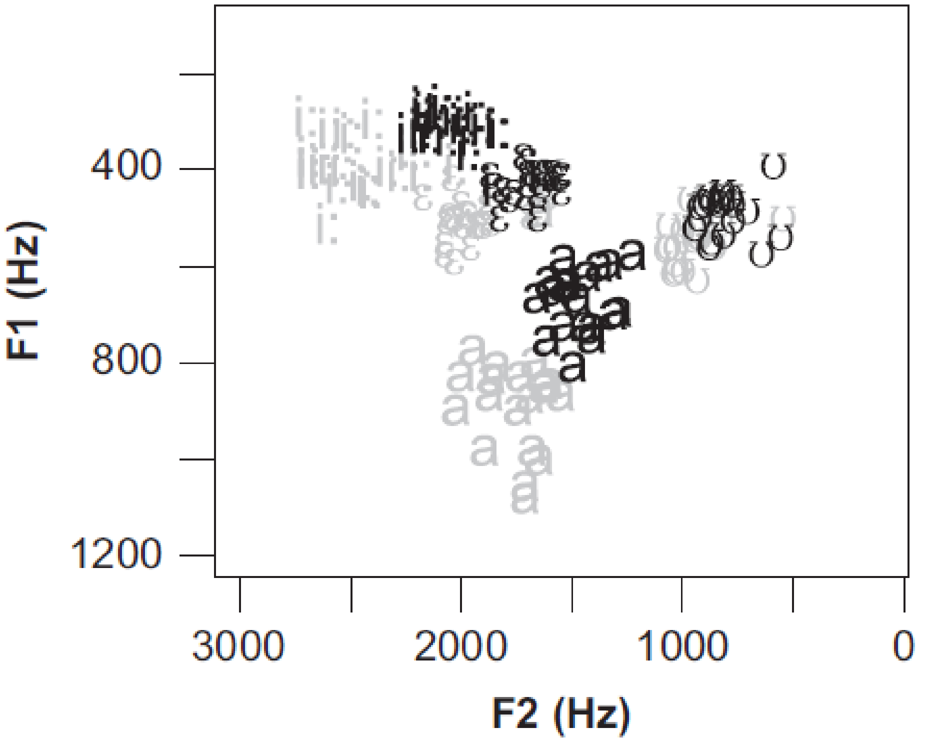


Figure 3.1: Acoustic vowel space for three vowels from male (black) and female (grey) speakers. Adapted from Weirich and Simpson (2014).

A speaker traversing a large acoustic vowel space in the same time as a speaker traversing a small acoustic vowel space is perceived as speaking faster, particularly in vowels with high F1 and F2 formant values (Weirich & Simpson, 2014). Since females typically have to traverse a greater acoustic space over the course of an utterance, listeners might be subject to the bias of tying perceived duration to acoustic complexity, even if measurable duration patterns are the same, or indeed point in the other direction (Weirich & Simpson, 2014).

3.1.2.2.2 Reported Differences in the Speech Rate of Male and Female Voices

Actual reported differences in the speech rate of males and females in the empirical literature is however somewhat quite different. Indeed, research tends to suggest that males do in fact have a faster speech rate than females. Byrd (1992, 1994) looked at the speech of 630 male and female American English speakers on several temporal parameters. They found that the mean sentence duration was 6.2% shorter in male speakers than in female speakers. Female speakers also had fewer vowel reductions (i.e., any change in the acoustic quality of vowels), all indicative of a faster speech rate in males than in female speakers. In line with this, Whiteside (1996) explored the speech of three male and three female English speakers. The results showed that females spoke with more pauses, longer sentence durations, and fewer vowel reductions and elisions (i.e., the omission of a sound, such as a vowel, from a word or phrase) than male speakers. Again, this indicates a slower speech rate for females. Pepiot (2014) analysed both dissyllabic and pseudo-words produced by 10 North-eastern American English speakers and 10 Parisian French speakers. The researchers showed that mean word duration was higher for female speakers in both languages (510ms for French speakers and 555ms for American speakers) than it was for male speakers (445ms for French speakers and 441ms for American speakers), confirming that speech rate was significantly faster for male speakers. Others have found that females produce longer sound durations than males, especially for vowel sounds, also suggesting that females have a slower speech rate than males (Simpson, 2009).

It has been suggested that any variation in the speech rate of male and female speakers may be related to differences in the expectations of male and female gender roles of a given society. Gender roles are socially constructed behaviours and attributes expected of individuals on the basis on being born either male or female (Basow, 1992). We learn appropriate gender roles in accordance with the expectations of a given society. In Western society, to be feminine is often thought to be nurturant, co-operative, and sensitive to the needs of others (Basow, 1992). To be masculine is to be more aggressive, dominant, and ambitious (Basow, 1992). Whilst people possess both sets of traits in varying degrees, social pressures and norms often lead individuals to conform to these expectations. Research has identified that several verbal indicators can affect perceptions of dominance and masculinity. For example, individuals who speak at a faster rate are often perceived as more dominant than those who speak more slowly (Aronvitch, 1976; Buller & Aune, 1988; Buller & Burgoon, 1986; Harrigan, Gramata, Lucic, & Margolis, 1989; Scherer, Lndon, & Wolf, 1973). Furthermore, Heffernan (2010) found that 'mumbling is macho', and that males tend to mumble more than females. Mumbling often results in sentences being spoken at a faster rate, presumably because the pronunciation of certain words are shortened considerably more than if they were spoken more clearly.

It is important to note that several experts have suggested that slower speaking styles in females could be the result of the laboratory testing conditions while reading aloud, making it difficult to explain any differences identified as a general basis for sex-related differences in speech rate. Nevertheless, whilst most studies do indeed investigate read speech (e.g., Byrd, 1992; Ericsdotter & Ericsson, 2001; Fitzsimmons, Sheahan, & Staunton, 2001; Pepiot, 2014; Simpson and Ericdotter, 2003; Whiteside 1996), comparable sex-specific durational differences have been found in both read and spontaneous speech, arguing against the suggestion that females speak slower than males only in laboratory conditions while reading aloud (Weirich & Simpson, 2014; Simpson, 2009). Despite this however, it is also important to recognise that whilst differences between the speech rate of male and females have been found, these differences can be small (Yuan, Cieri, & Liberman, 2006). What's more, several researchers have found no significant differences between male and female speech rate (Block & Killen, 1996; Robb, Maclagan, & Chen, 2004), and others have even found females to speak faster than males (Jacewicz & Fox, 2010).

In summary, reports of male and female difference in speech rate tend to suggest that males speak faster than females. However, research suggests that people believe females speak at a faster rate than males (Weirich & Simpson, 2014). Currently there is no clear consensus in the literature, and research on whether manipulations in speech rate are likely to change perceptions of speaker sex is considerably lacking. The findings reviewed suggest further exploration of speech rate as a cue to speaker sex is warranted.

3.1.3 Perceptions of Speaker Age

3.1.3.1 Fundamental Frequency (F0) and Speaker Age

3.1.3.1.1 Structural, Functional, and Hormonal Differences in the Fundamental Frequency (F0) of Male and Female Voices of Different Ages

Researchers generally agree that the F0 of an infant is around 500Hz and approximately 300 to 400 Hz above that which is ultimately achieved in adulthood (Aronson & Bless, 2011; Michelsson, Eklund, Leppanen & Lyytinen, 2002). By childhood (approximately age eight), this drops nearly 50% to around 250Hz for both males and females and is the result of the rapid growth of the larynx, vocal folds and surrounding support structures (Busby & Plant, 1995). The next marked drop in F0 is not seen until puberty, where the lowering of F0 is most likely caused by the structural and hormonal changes that occur. During this time, F0 reaches adult maturity and drops by another 50% in males to around 125Hz, but only 220 Hz in females (Titze, 1994). After puberty, mutational change of the voice is essentially complete and remains fairly stable into adulthood. F0 begins to change again from young adulthood into older age in both males and females. This is largely the result of anatomical changes in the larynx, however the pattern of this change is different in both males and females and is generally more noticeable in males (Linville, 1996). The onset of changes to the larynx is typically earlier in males than it is in females, starting in the mid 30's, and becoming most noticeable between the ages of 50 to 60 years.

As illustrated in Figure 3.2, the typical F0 pattern for male speakers follows a U-function, where F0 lowers from childhood to young adulthood and into middle age, and then rises again into older age. F0 for a male is typically lowest between 40 and 50 years, reaching the level of 20 to 30 years at age 60 to 70 years, and then continues to rise (Hollien & Shipp, 1972; Linville, 1996). The cause of the drop in F0 observed in young adult speakers into middle age is not fully understood, however it has been suggested that this continued decline may be due to subclinical trauma associated with normal voice use (Hollien & Ship, 1972). The increase in F0 is largely due to the anatomical changes in the larynx that occur with aging. Vocal folds become shorter, and there is increased stiffness in vocal fold tissues (Kadakia, 2013). These changes cause the vocal folds to vibrate more rapidly, increasing F0. Atrophy of the intrinsic muscles of the larynx is also typical. Degeneration of the thyroarytenoid (TA) muscles, cricoarytenoid (CA) muscles, and lateral cricoarytenoid (LCA) muscles also occurs, making it more difficult to lower F0 (Linville, 2001) (refer to Chapter 2, Section 2.3.2.1 for further details about the muscles in the larynx). The epithelium also thickens and the development of edema can occur. Edema is the swelling of the vocal folds due to accumulation of fluid in the superficial lamina propria (Linville, 2001). The resulting greater mass of the vocal folds interferes with normal vocal fold function by lowering F0 (Hirano, Kurita & Sakaguchi, 1989; Hixon et al., 2008). By approximately 85 years of age, a male's F0 will have risen to its highest level in adult life (Kreiman & Sidtis, 2013).

As illustrated in Figure 3.2, for females, the typical pattern of change is one where F0 continues to lower from childhood through to young adulthood, middle age, and older adulthood (Chatterjee, Halder, Bari, Kumar, & Roychoudhary, 2011; Benjamin, 1981; Ferrand, 2002; Linville, 2001; McGlone & Hollien, 1963). However, the most dramatic drop in F0 occurs at approximately 50 years of age (Beck, 1997; Chatterjee et al., 2011; Dehqan, Scherer, Dashti & Ansari-Moghaddam & Fanaie, 2012; Linville, 2001). The drastic drop in F0 is thought to be the result of hormonal changes that occur throughout the menopause (D’haeseleer, Vanlierde, Claeys & Depypere, 2012). During this time, there is a decrease in hormone production by the ovaries. Consequently, levels of estrogen and progesterone begin to fall. In the period immediately after the start of the menopause, the level of follicle stimulating hormone (FSH) and luteinizing hormone (LH) is very high, causing ovarian androgen production (Kadakia, Carlson, & Sataloff, 2013). These androgens deepen the voice and cause irreversible changes (Strauss, Mariah, Ligget & Lanese, 1985). After menopause, voices typically continue to drop in F0 because the ovary secretes little or no estrogen, but continues to secrete andrgoens (Sataloff, 2006). The vocal folds also thicken and the development of edema is likely to occur, both of which lower F0 (Chatterjee et al., 2011; Hixon, Weismer, & Hoit, 2008).

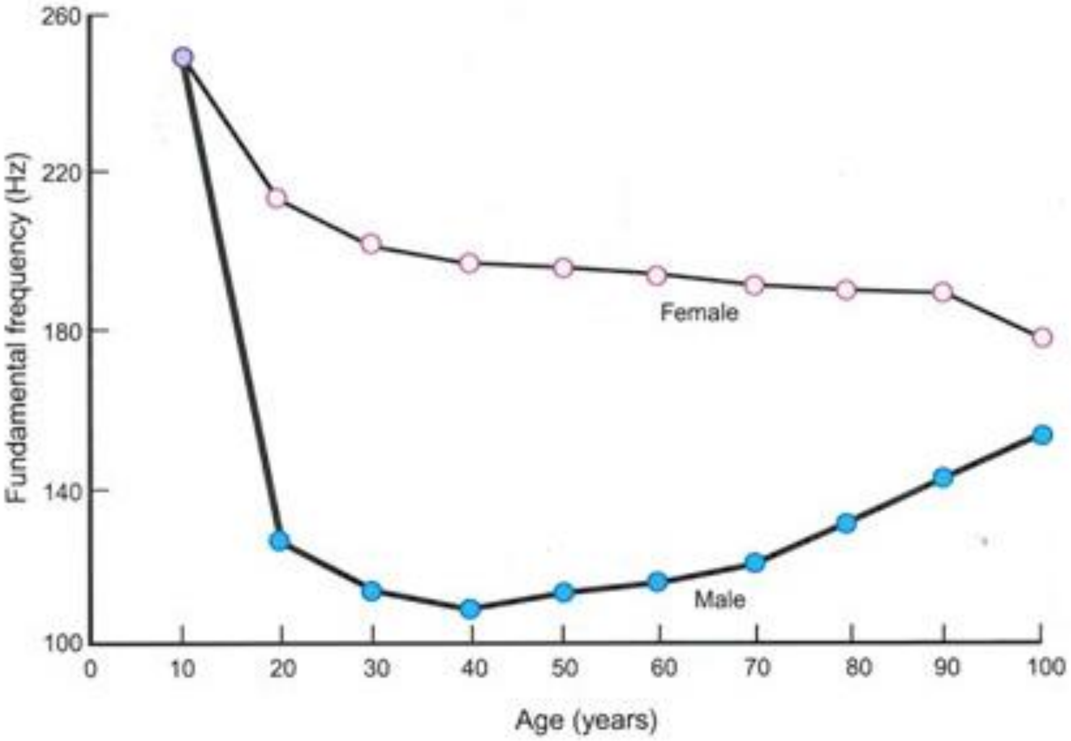


Figure 3.2: Fundamental Frequency (F0) changes from ages 10 to 100 years in males (blue circles) and females (pink circles). Purple circle indicates F0 for both males and females. Data obtained from various sources and rounded to the nearest decade for each sex.

3.1.3.2 Speech Rate and Speaker Age

3.1.3.2.1 Structural, Functional, and Behavioural Differences in the Speech Rate of Male and Female Voices

Speech rate is known to increase as a function of age in both males and females from childhood through to adulthood where it achieves its peak value around the mid 40's (Jacewicz & Fox, 2010; Kowal, O'Connell, Daniel, & Sabin, 1975; Walker, Archibald & Cherniakifish, 1992), before it begins to get progressively slower into older age (Bruckl & Sendlmeier, 2003; Harnserger, Shrivastav, Brown, Rotham & Hollien, 2008; Linville, 2001; Quene, 2008). The age-related increases in speech rate from childhood through to adulthood are not fully understood, however, they primarily appear to be because of gains in speech motor control abilities, and cognitive and linguistic processing (Goffman, Maassen & van Lieshout, 2010). Indeed, an essential process in the development of speech rate is the optimal tuning of the speech motor control system through motor learning (Nip & Green, 2013). However, immature control of the speech motor system (i.e., poor force and position control of the articulators) has been observed in young children and may contribute to slowed speech. Children have also been found to produce larger and slower articulator movements than adults (Goffman & Smith, 1999; Riely & Smith, 2003; Smith & Gartenberg, 1984; Smith & Goffman, 1998; Smith & Zelaznik, 2004; Walsh & Smith, 2002).

In terms of cognitive and linguistic processing, the relationship between speech rate and task demands suggest that children speak slower than adults because their articulatory movement speeds are slowed by the reduced capacity to formulate spoken language (Nip & Green, 2013). For example, children speak faster during simple speaking tasks, such as repetitions of simple syllables, than during more demanding tasks, such as conversational speech (Haselager, Slis, & Rietveld, 1991). Biological factors may also play a role, including anatomic growth, and neurologic and neuromuscular maturation (Maassen & van Lieshout, 2010). With regards to anatomic growth, children must develop articulatory performance stability as vocal structures undergo rapid changes in geometry and mass (Kent, 1984; Vorperian, Kent, Gentry, & Yandell, 1999). Changes in the size of the articulators will inevitably affect the co-ordination of speech (Callan, Kent, Guenther, Vorperian, 2000) and complicate the child's attempts to acquire a target acoustic output (Green & Nip, 2010). Furthermore, the speed at which an electrochemical impulse propagates down a neural pathway (i.e., conduction velocity) innervating orofacial structures have been found to increase with age (Barlow, Finan, Bradford, & Andreatta, 1993), suggesting that the slowed rate of speech in children may, in part, be due to the relatively slow conduction speeds in the central and peripheral nervous system.

Speech rate continues to get progressively slower into older age for both male and female speakers. The reduction in speech rate from adulthood to older age is primarily thought to reflect a slowing of the motor processes that occur due to changes in the supraglottic structures and articulators, changes in the structure and function of the nervous system, and a general weakening of the respiratory system. In terms of the supraglottic structures and articulators, aging causing muscle weakening and deterioration in motor control functions. Facial muscles begin to lose tone and elasticity, and atrophy of colleganous fibers occurs (Linville, 2001). Mechanisms controlling the articulators, including the jaw, tongue, lips, and soft palate also deteriorate making fine muscle co-ordination involved with speech production more difficult and reducing articulatory speed (Weismer & Liss, 1991). Specifically, the mandible (lower jaw) lengthens with age and becomes thinner due to bone resorption related to tooth loss (Israel, 1973). As a result, the points of attachment for masticatory (chewing) and facial muscles may be altered, reducing their biomechanical efficiency and speed capacity during speech (Kahane, 1981). Furthermore, the muscle strength of the tongue declines, reducing its mobility and range of motion (Rother, Wohlgemuth, Wolff, & Rebentrost, 2002). The reflex response of the lip muscles are also reduced (Hixon & Hoit, 2006), and thinning of the epithelium of the pharynx and soft palate occurs, making it increasingly difficult for accurate articulatory control.

Aging also changes the structure and function of the nervous system, including several that are important in speech motor control. For example, the primary motor cortex in the frontal lobe is the origin of the majority of descending axons to the motor neurons (Adams, 1987). Research has demonstrated changes in the neurons including irregular swelling and degeneration of dendrites, and fewer synapses (Adams, 1987; Scheibel, Tomiyasu, & Scheibel, 1977). Aging also results in the loss of dendrites in the cerebellum, the part of the brain which co-ordinates and regulates muscular activity (Willott, 1999). The number of motor neurons in the peripheral nervous system has also been found to decrease with age, with losses estimated as high as 1% per year beginning as early as the third decade of life, and accelerating after the age of 60 (Willott, 1999).

Reductions in respiratory power and breathing efficiency may also result in a slower speech rate. The most significant change is a loss in elasticity in lung tissue (Linville, 2004). Other respiratory system changes include stiffening of the thorax and weakening of the respiratory muscles, resulting in a loss in lung capacity, a decrease in maximum expiratory flow, and lung pressure (Huber & Spruill, 2008). Elderly speakers consequently experience a decline in the amount of air that can be moved into and out of the lungs (Linville, 2004), resulting in more frequent breaths needing to be taken (Hixon & Holt, 1987). Fatigue of respiratory and laryngeal structures, including the vocal folds and diaphragm, as a result of atrophy can also occur.

The differences in the rate at which elderly people speak compared to those who are younger or middle-aged could also be a result of the behavioural changes that occur. For example, older adults may speak slower in order to emulate a sociolinguistic pattern of speech typical of their age (Ramig, 1986). Consistent with these discrepancies between social expectation and actual voice characteristics is the finding that vocal portrayals of older adults typically use reduction in speech rate (Kreiman & Sidtis, 2010). Older adults may also adjust their speech to accommodate for the structural and functional changes that occur as a result of the aging process (Linville & Rens, 2001; Rastatter & Jacques, 1990). Indeed, Flethcer, McAuliffe, Lansford, and Liss (2015) provided some evidence that slower speech rate may be a behavioural strategy that older speakers implement so they are able to maintain articulatory precision.

3.1.3.3 Fundamental Frequency (F0) and Speech Rate as Cues to the Perceived Age of the Speaker

Research typically suggests that speech rate, and to a lesser extent F0, are important in estimating perceptions of speaker age (Waller & Eriksson, 2016). One way to study the effects of speech rate on age estimation is to ask listeners to make age estimates of voices from speakers who differ in chronological age. For example, Braun, Rietveld and van Bezooijen (1995) found that for male speakers, mean F0 does not influence perceptions of age, however, for speech rate, as the rate of speech gets progressively slower, perceived age increases. Others have found a positive correlation between mean F0 and perceived age for male speakers with an age range of 40 to 80 years (Horii & Ryan, 1981). Whilst this approach has been useful in identifying characteristics that may be important when making estimations of speaker age, associations can only be made. Furthermore, such research has typically focused on male speakers only making it difficult to determine whether the same characteristics are used when making estimations of age for female speakers too.

Others have asked listeners to attribute vocal characteristics to different groups of speakers. For example, Ptacek and Sander (1966) asked listeners to attribute characteristics to male and female voices of younger adults (under 35 years of age) and older adults (over 65 years of age). The results showed that speakers who were classified as older had significantly slower speech rates than the younger adult group. In other work, Shipp, Qi, Huntley, and Hollien (1992) recorded samples of continuous speech from males from three different age groups (27-35, 53-57, and 75-85 years of age), and asked listeners to attribute vocal characteristics that were typical to each group. Both F0 and speech rate were found to be good predictors of speaker age. Young speakers were perceived as having faster speech rates than middle-aged and old speakers. Listeners also associated a low F0 with old age speakers. However, differences in F0 between young and middle-aged were not found to be statistically significant. This suggests that F0 may not be a reliable enough cue on its own to determine speaker age, or that acoustic information other than F0 may be used to distinguish the age of male speakers in this range. In another study, Ryan & Burk (1974) asked listeners to determine the presence or absence of several vocal characteristics of 80 male speakers between the ages of 40 to 80 years old. Results indicated that a slow rate of articulation was a strong predictor of perceived age, with younger adult males being attributed a faster speaking rate compared to older adult males. Of course, the findings from such studies do not establish whether the same acoustic characteristics are used when making estimations of age for female speakers.

To determine the cues to perceived age in female voices, Linville & Fisher (1985) asked female listeners to judge the age of both whispered and phonated vowels produced by 75 females aged between 25-35, 45-55, and 70-80 years. Overall, voices with a lower F0 sounded older. Listeners also had significantly higher accuracy rates when judging age from phonated vowels as opposed to whispered vowels, suggesting that F0 is a powerful and resilient cue to perceived age in female speakers. For whispered vowels (where mean F0 was absent), speakers perceived as old had lower F1 formant values than did other speakers. No such correlation was observed in normally phonated speech, suggesting that listeners appear to ignore resonance cues that are available in the acoustic signal if F0 cues are also available. Nevertheless, the contribution of speech rate to estimates of speaker age was not established. Furthermore, the study employed only female listeners making it difficult to determine whether male listeners also use F0 when making age estimations for female speakers.

Few studies have made use of both male and female speakers and listeners in their work. Hartman and Danhauer (1976) informed a set of male and female listeners of the mean perceived age of male and female speakers and asked them to write down descriptions of the voices. Older speakers were rated as having a slower speech rate than younger speakers. However, changes in speech rate were perceived as occurring at a much younger age than they actually do, suggesting that discrepancies may exist between listener's expectations about speakers of different ages and the vocal characteristics that actually exist. Listeners were also found to consistently associate lower F0 with old age in both male and female speakers despite reported increases in F0 with age in males, suggesting the presence of vocal stereotyping by listeners regarding F0 and speech rate with age (Hartman & Danhauer, 1976).

It is often acknowledged that experimental work in which the parameter of interest is manipulated constitutes much harder causal evidence for the effects of acoustic cues on age estimations (Waller, Eriksson, & Sorqvist, 2015). To date however, very few studies have investigated the effect of F0 or speech rate on perceived age using this method. Shrivastav, Hollien, Brown, Rothman, and Harnsberger (2003) resynthesized 16 natural male voices of young (20 to 33 years) and old males (aged 70 to 90 years) in F0 and speech. The voices of older speakers were decreased in F0 and increased in speech rate (to make them sound younger), whereas the voices of younger speakers were increased in F0 and decreased in speech rate (to make them sound older). A significant shift in age estimates were observed for the older, but not younger, speakers in manipulations of F0 or speech rate. For speech rate, estimates of perceived age were lower (i.e., younger) when speech rate was increased. For F0, estimates of perceived age were lower (i.e., younger) when F0 was decreased. This finding is particularly surprising for female voices given that F0 typically decreases in female speakers as age increases. The results therefore lend further support to the suggestion that some stereotyping of the vocal characteristics for female speakers may exist. Moreover, the effects of the manipulations were greater in magnitude for older speakers in comparison to younger speakers, suggesting that speech rate and F0 may gain greater importance as perceptual age cues with increased speaker age. Similar findings have also been found for male and female voices when manipulations in F0 or speech rate are made for sentences (Harnsberger, Shrivastav, Brown, Rothman, and Hollien, 2008), single words (Winkler, 2007), and isolated vowels (Smith, Walters, & Patterson, 2007). Smith & Patterson (2005) did find manipulations in F0 to affect perceptions of speaker age, however, they only used male voices and so the findings are difficult to generalise to female voices. Nevertheless, taken together, the results tend to suggest that estimates of speaker age may be influenced by manipulations in speech rate more strongly than manipulations in F0.

Waller and Eriksson (2016) observed the effects of spontaneously manipulating F0 or speech rate on perceptions of speaker age. In the first part of their work, male and female speakers in different age groups (20 to 25 years, 40 to 45 years, and 60 to 65 years) read a short text under three voice conditions. In the first condition they used their natural voice, in the second condition they attempted to sound 20 years younger, and in the third condition they attempted to sound 20 years older. The researchers identified that speakers increased F0 and speech rate when attempting to sound younger and decreased F0 and speech rate when attempting to sound older. This strategy was applied regardless of speaker sex or age, suggesting that the speakers modified their voices according to their stereotypes of how young and old voices sound. In the second part of their work, participants listened to speech samples from the three voice conditions listed above and estimated the age of the speakers. The results suggested that the manipulations were effective in that the voices in the manipulated conditions received age estimates in the attempted direction (i.e., speakers attempting to sound older were estimated as older, and speakers attempting to sound younger were estimated as younger), although the changes in age estimates were small. This finding was held for both male and female voices and there was no difference in the effectiveness between attempts to sound younger and to sound older. When listeners were asked what cues they used to make estimations of age, results indicated that listeners use speech rate, but not F0, as a cue to speaker age. The findings therefore provide further support to the importance of speech rate as a cue to perceived age. However, this makes it particularly difficult to determine the relevant contributions of F0 or speech rate on estimates of speaker age as it is impossible to control for changes in one cue whilst manipulating the other.

In summary, it appears that the decline in overall speech rate may be the most important indicator of perceived aging in the voice. However, F0 may also be used to make estimates of speaker age. Estimations of speaker age appear to be influenced by the stereotyping of vocal cues rather than the actual changes that occur over the lifespan. Nevertheless, the literature examining the relationship between perceptions of speaker age and the acoustic cues of the voice is still surprisingly sparse, and several methodological issues make it difficult for any clear conclusions to be drawn.

3.2 Recognition Memory for Voices

Section 3.1 considered the effect of manipulations in F0 or speech rate on perceptions of the speaker’s identity, sex, and age. Manipulations in F0 or speech rate may also be important when trying to recognise the voice. Research into the factors that affect recognition for a speaker has a long history, dating back over 50 years. For example, studies have considered the importance of speaker variables including ethnicity, other race, and accented voices (e.g., Kerstholt, Jansen, Van Amelsvoort, & Broeders, 2006; Lass, Mertz, & Kimmel, 1978; Phillipon, Cherryman, Bull, & Vrij, 2007), familiarity (e.g., Abberton & Fourcin, 1978; Hollien, Bennett, & Gelfer, 1983; LaRiviere, 1972), disguise (e.g., Orchard & Yarmey, 1995; Saslove & Yarmey, 1980), and emotional stress and arousal (e.g., Hollien & Majewski, 1977; Reid & Craik, 1995; Saslove & Yarmey, 1980). Others have studied listener variables which include age (e.g., Hashtroudi & Ferguson, 1995; Hashtroudi, Johnson, & Chrosniak, 1989; Yarmey, 2000), sex (e.g., Roebuck & Wilding, 1993; Wilding & Cook, 2000; Yarmey, 1986; Yarmey & Matthys, 1992; Yarmey et al., 2001), blindness (e.g., Cobb, Lawrence, & Nelson, 1979; Muchnik, Efrati, Nemeth, Malin, & Hildesscheimer, 1991; Winograd, Kerr, & Spence, 1984), training and skill (Cobb, Lawrence, & Nelson, 1979), confidence (e.g., Olssen et al., 1998; Orchard & Yarmey, 1992; Yarmey & Matthys, 1992), and emotional stress and arousal (e.g., Read & Craik, 1995). Research has also emphasised the role of situational variables including the content of the spoken message (e.g., Reid & Craik, 1995), the duration of the speech sample (e.g., Bull & Clifford, 1984; Clifford, 1980; Cook & Wilding, 1997; Orchard & Yarmey, 1995), and the length of retention interval (e.g., Clifford, Rathborn, & Bull, 1981; Kerstholt, Jansen, van Amelsvoort, & Broeders, 2004; 2006; van Wallendael, Surace, Parson, & Brown, 1994; Yarmey & Matthys, 1992). Such work has typically used identification tasks which involve comparing a voice the listener has just heard to an exemplar or representation stored in memory (Kreiman & Sidtis, 2013). Recognition occurs once listeners determine that the voice they have heard is the one the voice that they previously heard, whether or not they are able to name the speaker.

Of the many factors that have been investigated, very few have considered the impact of manipulations in the acoustic cues of the voice and how they affect recognition performance.

This is important because intra-speaker variation in a speaker's voice (whether unintentional or deliberate) exists, and can greatly reduce recognition performance (Reich & Duke, 1979). Furthermore, people who are asked to recognise a speaker from their voice are likely to be faced with such difficulties. Understanding what cues are likely to be accurately remembered, and those that are subject to distortion, also has substantial applied interest. During a criminal investigation, it is likely that the police will ask the victim or witnesses of a crime to identify the suspect from a voice recording, particularly in situations where the suspect is encountered under poor visual conditions (Yarmey, 2001; 2004). It is important for law enforcers to be aware of the errors that can occur as a result of the intra-individual variations that exist in a speaker's voice in order to enhance the accuracy and relevance of testimony in court.

The studies that do exist on this topic have set out to determine whether memory construction processes produce distortions for representations of acoustic cues in memory and whether these distortions are predictable. The findings have suggested that categorical memory processes distort voice memory for F0 (Mullenix et al., 2010; Stern et al., 2007) and for speech rate (Mullenix et al., 2010) in a manner where memory is exaggerated. Whilst insightful, the findings are limited, and there are several methodological issues that make it difficult for any clear conclusions to be drawn. The following section will address the extent to which manipulations in F0 or speech rate can lead to errors in memory for these acoustic cues of the voice. The review will begin with a discussion of the research that has considered memory categorisation processes and the mistakes that arise in memory for F0 and speech rate because of this. It will then move on to consider other factors that might contribute to performance on a memory task for voice F0 and speech rate. It discusses the time course of echoic memory, and reviews research that has found people to be more accurate in a task that uses shorter intervals between presentations of the stimuli. It also discusses how memory for the voice may be somewhat easier if the sentence spoken is the same throughout the duration of the task, and it will consider the possibility that any biasing affecting performance may be dependent on these factors.

3.2.1 Memory Categorisation and The Accentuation Effect

To function efficiently in the social world, we must quickly make sense of our multifarious and fast-changing environment (Brosch, Pourtois, & Sander, 2010). However, human cognitive processing resources are limited and this presents a challenge in a rapidly changing social environment. Given these limitations, people devise short-cut strategies to simplify the nature of incoming information. One proposed strategy is categorisation in which it is assumed that stimuli are reduced into cognitively simple categories which contain other stimuli that are equivalent/analogous to each other (e.g., same colour, same shape, same tone) and different from other stimuli (Brosch, Pourtois, & Sander, 2010). Categorisation is an important cognitive process (Gifford, Cohen, & Stocker, 2014). The process of categorisation means that it becomes less cognitively effortful when an observer encounters a new stimulus. However, the act of placing stimuli into distinct categories can lead to distortions which result in the stereotyping of some distinctive features (Hogg & Vaughan, 2010). For example, when stimuli covary by constant amounts on a given continuum, people are less likely to perceive stimuli within the same category to be different than when stimuli are placed in different categories. In other words, people minimise the perception of differences within a category and maximise the perception of differences between categories. Consequently, when people are asked to recall properties of stimuli within a category, they tend to recall features typical of the category overall, rather than the individual properties of the stimulus. This is known as the *accentuation effect* (Fiske, Gilbert, & Lindzey, 2010; Huart, Corneille, & Becquart, 2005; Sutton & Douglas, 2013).

3.2.1.1 Accentuation Effects in Non-Social and Social Stimuli

Accentuation effects have been found to be real and robust and have been observed with both non-social and social stimuli. In their seminal work, Tajfel & Wilkes (1963) demonstrated how the placement of a category boundary between lines of varying length caused the lines in the long category to be judged as longer and the lines in the short category to be judged as shorter than when no category was provided (Tajfel & Wilkes, 1963). Others have shown that people typically overestimate temperature variations between different months of the year compared to temperature estimates within the same month (Krueger & Clement, 1994), and that objects belonging to categories with redder objects are judged as more red than identically coloured objects belonging to different coloured categories (Goldstone, 1995). In terms of social stimuli, ratings of statements attributed to the same newspaper have been found to be judged more similarly than those from different newspapers (Eiser, 1971). People have also been found to describe a person as having a greater shared identity to themselves when that person has a stronger assimilation to the participant's own position (Haslam & Turner, 1992), and to judge a person's personality as being more similar to another's when they are placed in the same group (Krueger & Rothbart, 1990; Queller, Schell, & Mason, 2006).

More recent work has shown how accentuation effects can also affect perceptions of facial stimuli. For example, adding a featural characteristic of a particular race (such as a Hispanic or African American hairstyle) to a facial composite leads people to judge faces as more typical of that racial origin compared to when no modification or labels were used (MacLin & Malpass, 2001). Similar results have been observed in other studies where faces have been given a more white European name (Hilliar & Kemp, 2008), or if the faces have been labelled as 'black' (Levin & Banaji, 2006). Others have shown that categorising faces can lead to errors in memory at the recognition stage. For example, morphed faces possessing more or less stereotypical features of a particular race were misremembered as being more prototypical of that particular race than they actually were (Corneille, Huart, Becquart, & Brédart, 2004). Comparable effects have been found when using gender ambiguous faces (Huart, Corneille, & Becquart, 2005), and ambiguous angry and happy faces (Halberstadat & Niedenthal, 2001).

3.2.1.2 Accentuation Effects in Memory for Voices

Surprisingly, very few researchers have considered categorisation or accentuation effects in relation to voices. This is remarkable because variations in the paralinguistic characteristics of the voice can occur within the same speaker (within-speaker, or intra-speaker variation). Speakers rarely pronounce given words or phrases in an identical way on different occasions, even if the second utterance is produced in close succession (Hollien, 1990). The same speaker can also sound different from time-to-time because of factors such as time of day, fatigue, intoxication (from alcohol or drugs), thought distractions, situational demands, mood state, changes in health and physical status, stress, and a speaker’s emotional state (Nolan, 2005; Saslove & Yarmey, 1980). Speakers can also modify their own voice by means of disguise (refer to Chapter 2, Section 2.3.1). Whilst listeners are often robust to these changes and have little difficulty identifying speakers using their voice alone, researchers have stressed how such changes can introduce great acoustic variation and increase recognition errors (Endres, Bambach, & Flosser, 1971; Reich, Moll, & Curtis, 1976). Furthermore, it is possible that listeners categorise voices in terms of their acoustic properties, which might then lead to errors when attempting to recognise these at a later date. Indeed, studies have shown that we attend to acoustic properties of a sound to make categorical judgements (Marcell, Barello, Greene, Kerr, & Rogers, 2000), and that we categorise speech sounds according to their frequency (i.e., high, moderate, and low frequency) (Mondor, Hurlburt, & Thorne, 2003; Wong, 1976; Xu, Krishnan, & Gandour, 2006).

Mullenix et al. (2010), one of the few studies to explore this topic in voices, found evidence for accentuation effects for voice memory. The researchers investigated the effects of manipulating fundamental frequency (F0) and speech rate (words per minute) on recognition memory for voices. To do this, Mullenix et al. (2010) created a number of versions of a male synthesised target voice; a version that was higher than the original voice and fell within the higher F0 speaking range (which they labelled ‘high F0’), a version that was lower than the original voice and fell within the lower F0 speaking range (labelled ‘low F0’), and the original version of the voice which fell in the moderate F0 speaking range (labelled ‘moderate F0’). Similar manipulations were also applied for the speech rate condition to obtain target voices that were faster in rate (labelled ‘fast rate’), slower in rate (labelled ‘slow rate’), and the original version (labelled ‘moderate rate’). This resulted in six conditions of interest (i.e., high, moderate, and low F0, and fast, moderate, and slow speech rate). Using a two-alternative forced choice (2AFC) voice recognition task, participants were presented with one of the target voices and were then asked to recognise this from a pair of sequentially presented voices. The paired voices included the previously heard target voice and a distractor voice which consisted of a modulated version of the target (which was either higher or lower in F0, or faster or slower in speech rate). The results showed a predictable pattern of memory errors. Listeners mistakenly selected voices lower in F0 than the low F0 target voice, and voices higher in F0 than the high F0 target voice. However, there was no difference in the selection of higher or lower F0 distractor voices for moderate F0 target voices. In contrast, for speech rate, listeners mistakenly selected voices slower in rate than the slow rate target voice. However, there was no difference in the selection of faster and slower rate distractor voices for moderate and fast rate target voices. The results are shown in Figure 3.3.

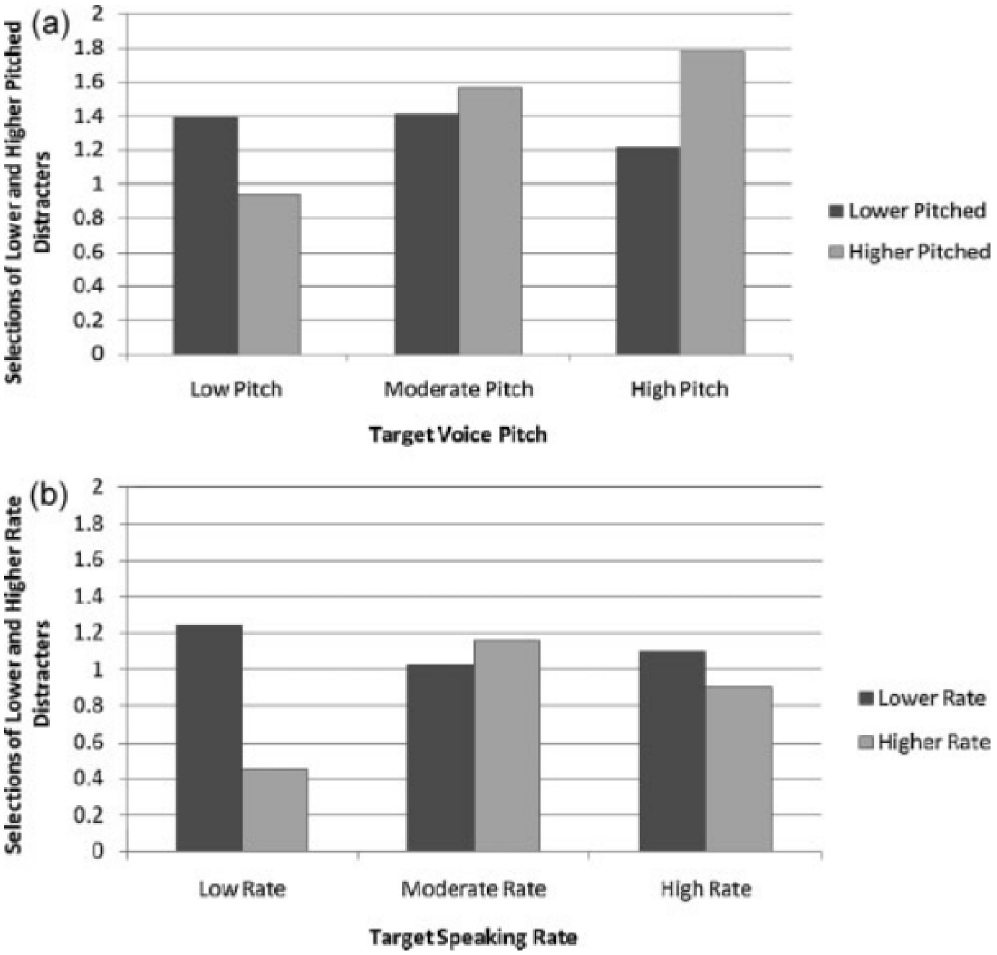


Figure 3.3: Figure obtained from Mullenix et al., (2010). Panel (a) shows the mean number of errors made (i.e., selection of distractor voices higher or lower in F0) for the three target F0 voice conditions. Panel (b) shows the mean number of errors made (i.e., selection of distractor voices faster or slower in speech rate) for the three target speech rate voice conditions.

According to Mullenix et al. (2010), the effect of increased recognition errors in the low and high F0 conditions likely reflects an accentuation effect. They argue that listeners place the higher and lower F0 voices they hear into cognitively simple categories, leading them to recall features most salient to that category (i.e., a higher or lower F0) rather than the individual properties the voices actually have. A similar pattern of findings has also been found for F0 using both a male and female synthesised voice (Stern et al., 2007). The absence of an effect for speech rate is not unexpected since, unlike F0 which under normal circumstances is relatively stable, within-speaker variation in speech rate can be highly variable; sometimes people speak quickly, while other times they speak slowly. Thus, it is likely that listeners are more familiar with speech rate variability and hence, are more robust to the changes that occur as a result of the manipulations made. As a consequence, different properties of the voice may be more or less susceptible to category-based memory distortions. Listeners may be better able to recognise a voice when changes to speech rate are made compared to changes in F0. However, given the limited number of studies that have considered accentuation effects in relation to voices, it is difficult for any clear conclusions to be drawn.

3.2.2 Other Factors Contributing to Recognition Memory for Voices

3.2.2.1 Increasing the Inter-Stimulus Interval

3.2.2.1.1 The Echoic Memory Store

In line with Atkinson and Shiffrin’s (1968) modal model of human memory, information processing occurs in a series of stages consisting of sensory memory, short term memory (STM; or working memory), and long term memory (LTM). Sensory memory is the briefest element of human memory and refers to the ability to retain impressions of sensory information after the original stimuli have ended. Sensory memory is thought to contain separate modality-specific storage systems for each sensory channel (Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974; Baldwin, 2012). This means that each sensory register retains information specific to certain sensory information (Radvansky, 2012). Echoic memory is a component of sensory memory that is specific to temporarily retaining auditory information. Echoic memory is thought to hold an exact replica (in the form of an auditory trace or echo) of the information presented (Baldwin, 2016). Precise representations of auditory information in sensory memory degrade quickly. The echoic memory store retains information for approximately 3 to 5 seconds (Glanzer & Cunitz, 1966; Lu, Williamson & Kaufman, 1992; Treisman, 1964) and decays exponentially over time (Lu, Williamson, & Kaufman, 1992).

Once the sensory memory trace has decayed or is replaced by a new memory, the information stored is no longer accessible and ultimately lost.

3.2.2.1.2 Auditory Sensory Memory and Recognition Performance

Performance on a memory task for auditory stimuli may be dependent on the time course of the echoic memory store. The more precise the mental representation of the auditory stimulus, the more accurate the memory for that stimulus is likely to be. Retrieval from echoic memory is likely to be relatively easier at shorter intervals because the acoustic trace is stronger. Therefore, we might expect people to be more accurate in a task that uses shorter inter-stimulus intervals between presentations of auditory stimuli than on one where the interval is longer in duration. Indeed, several authors emphasise the importance of timings in the range of tens of milliseconds to a few seconds in auditory sensory tasks (Ivry & Spencer, 2004; Mauk & Buonomano, 2004; Buhusi & Meck, 2005; van Wassenhove, 2009). Research tends to suggest that the ability to detect changes in tones deteriorates as the inter-stimulus interval increases. For example, in one experiment, comparisons were made using a fixed standard tone and a comparison tone while varying the inter-stimulus interval (Harris, 1952). Results indicated an appreciably greater decline in discrimination with the comparison tone as the inter-stimulus interval increased above 3 seconds. Performance on a same-different task has also been found to steadily decrease when a variable delay of 0.5 to 2 seconds was placed between two tones (Kinchla, 1973). In other work, Wickelgren (1969) compared recognition memory for frequency of a standard tone and a comparison tone separated by variable delay intervals (0 to 180 seconds). The decay of the memory trace was found to occur temporally over time, and regardless of the frequency difference between the two tones.

Short inter-stimulus intervals between presentations in voice recognition tasks may facilitate comparisons between voice stimuli, thereby increasing accuracy. This is because presenting the to-be-compared stimuli within a short time frame likely facilitates appraisals based on high-quality voice representations (Smith, Dunn, Baguley, & Stacey, 2016). Most previous tests of voice recognition accuracy have presented voices close together in time, with a standard inter-stimulus interval of either 500ms or 1 second (e.g., Lachs & Pisoni, 2004; Mavica & Barenholtz, 2013; Mullenix et al., 2010; Stern et al., 2007). Whilst insightful, such studies do not consider this period of decay in memory. Thus, performance rates are likely to be misrepresented. Several studies have explored the effects of delay on recognition memory over several days or weeks (e.g., Clifford, 1983; Goldinger, 2004; Kowalska, 1997; Palmer, Havelka, & Hooff, 2013). Nevertheless, such tests only inform us of the role of decay in long term memory. What is more, the length of the delay means that any effect of extraneous variables on memory during that period cannot be controlled for, making it particularly difficult to determine whether performance is purely a result of decayed memory. Increasing the inter-stimulus interval by a few seconds may therefore give us a more precise estimate of the rate of decay for voices in auditory memory.

To this authors knowledge, no research has explored whether increasing the inter-stimulus interval by a matter of seconds can affect performance on a voice recognition test. However, the results of several studies tend to suggest that accuracy for speech stimuli may deteriorate quickly. For example, Pisoni (1973) used a same-different speech discrimination task to determine whether two vowel sounds were identical physically or not. The time delay between the two vowel stimuli were set at intervals of 0.5 or 2 seconds. Performance was significantly poorer when the vowel sounds were separated by longer separations (i.e., 2 seconds). Crowder (1982) conducted two experiments on same-different vowel distinctions. Inter-stimulus intervals of 0, 2, 4, 6, 8, 10, 12, 14, 16, and 18 seconds were made. Longer delays led to significantly poorer discrimination than shorter delays. However, the auditory memory loss appeared asymptotic at about 3 seconds. Hanson (1977) also found poorer performance in a same-different task using spoken syllables with an inter-stimulus interval of 2.5 or 5.7 seconds. Listeners were found to be less accurate when comparing stimuli at 5.7 seconds compared to 2.5 seconds. The above studies indicate that judgements may depend, at least partly, on auditory sensory memory - the auditory memory trace of the first stimulus was decaying during the inter-stimulus interval, thus making it more difficult to detect differences between them.

3.2.2.1.3 Auditory Sensory Memory and the Accentuation Effect

Any bias affecting performance may also be dependent on this time course. In line with this suggestion, research has shown that when event details fade from memory over time, people are more likely to rely on schematic information to complete (or embellish) those faded memories, resulting in an increase in stereotype-consistent errors (e.g., Greenberg, Westcott, & Bailey, 1998; Kleider, Pezdek, Goldinger, & Kirk, 2008; Lampinen, Faries, Neuschatz, & Toglia, 2000; Neuschatz, Lampinen, Preston, Hawkins, & Toglia, 2002). As noted in Section 3.2.1.2, memory for F0 and speech rate of the voice has been found to reflect category typical representations rather than the specific features of items (Mullenix et al., 2010; Stern et al., 2007). Such studies used a 1-second inter-stimulus interval between presentations of the voice stimuli. The task may be relatively easy because the acoustic trace is likely to be strong. It is quite possible that as the inter-stimulus interval increases and the task becomes more difficult, listeners become increasingly reliant on category based information stored in memory to aid recognition. We might therefore expect more errors to be made when matching a voice to a previously heard target voice. To date however, no research has explored this idea further with voice stimuli.

3.2.2.2 Changing the Spoken Message

3.2.2.2.1 Principles of Pattern Recognition

The principles of pattern recognition help to explain how we recognise, identify, and categorise incoming sensory information from the external world. Specifically, this cognitive process refers to the ability to match information from a stimulus with information stored in memory (Eysenck & Keane, 2000). By comparing the information to a variety of stored candidates, humans are better able to recognise a stimulus by matching the one that it most closely resembles in memory. Pattern recognition relies on both bottom-up and top-down processing; the stimulus information arrives from the sensory receptors (bottom-up processing), and the incoming information is matched to patterns that already exist in memory derived from people's knowledge and previous experiences (top-down processing). Pattern recognition is fundamental to numerous aspects of human cognition. Among many, recognised patterns can be those perceived in facial features (e.g., Vernet, Martin, Baudouin, Tiberghien, & Franck, 2007), units of music (Krumhansl, 2001), objects (e.g., Schneiderman & Kanade, 1998), components of language (Margolis, 1996), or characters and other symbols (Eysenck & Keane, 2003).

3.2.2.2.2 Changing the Sentence and Recognition Performance

Based on the principles of pattern recognition, it could be presumed that recognition of the voice may be somewhat easier if the sentence spoken is the same as the one that was previously heard. Recognition of the voice may be achieved by mapping the auditory information of the spoken sentence onto stored representations in memory (Weber & Scharenborg, 2012). By repeating the same sentence, listeners can use patterns identified from the spoken sentence to determine whether the voice heard matches the mental representation of the voice stored in memory. Consequently, memory for the voice might be achieved without any knowledge of the voice per se. Rather, recognition might be accomplished on the basis of a simple familiarity judgement, i.e., ‘does this particular pattern match that of what I previously heard?’ (Glisky, Rubin, & Davidson, 2001). However, if someone were to speak a different sentence to one previously heard, recognition of the voice would require retrieval of the previously spoken sentence, some comparison of the two sentences, and information about the voice linking the two together (Glisky, Rubin, & Davidson, 2001). Although a decision might still be made on the basis of familiarity, the judgment is almost certainly more difficult as it requires knowledge about the voice itself.

Studies have shown that recognition performance is superior when identical test sentences are used. For example, Reid and Craik (1995) examined the effect of voice passages on recognition memory over a time delay of 17 days and found that recognition performance was better when listeners heard an original recording compared to when they heard a different passage. Using an old/new recognition test, Winograd, Kerr, & Spence (1984) found that memory was improved when a voice repeated itself relative to when it was saying something new. They suggested that voice recognition is largely dependent on recognising the features that are distinctive to that particular voice. By this argument, a voice reading a repetition of the same message is more likely to reproduce such distinctive features than when it is reading a new message. For example, if a speaker has an unusually sibilant /s/ (spoken in the study phase), yet there is no /s/ in the sentence at the test phase, then recognition is likely to be hindered because the listener is unable to match this particular feature to the representation stored in memory. And feature overlap is maximal, of course, when the same message is repeated (Winograd, Kerr, & Spence, 1984). Glisky, Rubin, & Davidson (2001) recorded several male and female voices speaking two sentences that were equated for speaking time. Using a 2AFC task, results showed that voices speaking the same sentence at study and test were more likely to be identified correctly than voices speaking a different sentence. The researchers concluded that information about the conjunction of the voice and the sentence were encoded at the study phase and that this may serve to enhance memory for the voice when it is heard again at test speaking the same sentence. By repeating presentations of the voice speaking the same sentence, components of the voice that are involved in the specific voice-sentence pairing can be further strengthened and compared to the representation stored in memory.

Several studies (e.g., Legge, Grosmann, & Pieper, 1984; Nygaard & Pisoni, 1988; Sheffert, Pisoni, Fellowes, & Remez, 2002; Shefferet & Olson, 2004; Zaske, Volberg, Kovacs, & Schweinberger, 2014) have also found recognition performance to be high for voices tested with previously unheard speech samples. Some have proposed that these findings are indicative that humans acquire representations in memory that store idiosyncratic voice properties, and thus allowing voice recognition to occur independent of speech content (Zaske, Volberg, Kovacs, & Schweinberger, 2014). However, it should be noted that these studies used voice learning where the same sentence was repeated over many exposures (sometimes over a number of days – e.g., Sheffert, Pisoni, Fellowes, & Remez, 2002), before a different sentence was used at the testing stage. It is possible that listeners became increasingly familiar with the voices over repeated presentations. Therefore, performance rates are likely to have been higher than if they had only heard the voice once before. Indeed, research has shown that listeners can recognise familiar voices from variable utterances even in the first instance (Skuk & Schweinberger, 2013). What is more, recognising familiar and unfamiliar voices have been found to be separate functions (Van Lancker & Kreiman, 1986), suggesting that the approach used was not appropriate for testing recognition memory for unfamiliar voices.

In summary, the research tends to suggest that for unfamiliar voices, decisions as to who is speaking are likely to be made on the basis of pattern matching and familiarity judgements. This might reflect a degree of inter-dependence between speech (i.e., linguistic information) and voice (i.e., non-linguistic information) in speaker recognition. Therefore, unfamiliar voice recognition is likely to be a speech-dependent, as opposed to a speech-invariant, process (Glisky, Rubin, & Davidson, 2001). Studies using the same sentence spoken at the study and testing stage may therefore underestimate recognition errors for unfamiliar voices. This has important implications for earwitness memory. Several experts (e.g., Broeders & Rietveld, 1995; Bull & Clifford, 1999; Hammersley & Read, 1996; Hollien, 1996; 2002; Hollien & Huntley, 1995; Ormerod, 2001) have suggested a number of criteria for voice lineups, including the use of a non-identical speech phrase in order to prevent the deliberate distortion of specific words or phrases by a guilty suspect at the time of recording. However, given that recognition performance is superior for identical test sentences, earwitness memory may actually be impeded if a different phrase is used at the identification stage.

3.2.2.2.3 Changing the Sentence and the Accentuation Effect

Any biasing affecting performance may also be dependent on the spoken message. Memory for F0 and speech rate have been found to reflect category typical representations rather than the specific features of the voice when identical sentences are used at both the study and testing phase (Mullenix et al., 2010; Stern et al., 2007). However, given the likely interdependence between linguistic and non-linguistic information in unfamiliar speaker recognition, using the same sentence might have assisted listeners in the recognition process. This is because listeners would have been able to make use of patterns identified from the spoken sentence to determine whether the voice heard matches the mental representation of the voice stored in memory. Thus, any errors made may be more pronounced when a different sentence is used to the sentence that was previously heard because it is more difficult for listeners to make judgements about the voice based on linguistic information alone. Consequently, listeners are likely to become increasingly reliant on paralinguistic properties (i.e., F0 and speech rate cues) of the voice. Category typical representations stored in memory may be used more when making decisions about the voice, resulting in a further biasing of the characteristics properties of the voice. We might therefore expect a further increase in accentuation errors when matching a voice to a previously heard target voice. To date however, no research has explored this idea further.

3.3 Summary Conclusions

In light of the findings from the present review, there is evidence for perceptual links between acoustic cues of the voice and characteristics of the speaker. It is apparent that manipulations in F0 are particularly important when determining the identity of the speaker (e.g., Kuwabara & Takagi, 1991; Lavner, Gath, & Rosenhouse, 2000; Sell, Suied, Elhilali, & Shamma, 2015). The evidence for speech rate is limited. However, the results so far suggest that manipulations in this cue may be of use when attempting to determine the identity of the speaker (e.g., Brown, 1981). Manipulations in F0 may also be important when making judgements about the sex of the speaker. In the main, the evidence tends to suggest a male-advantage in the perception of speaker sex for both male and female voices (e.g., Assman et al., 2006; Coleman, 1976; Gelfer & Mikos, 2005). However, the evidence is not always clear cut, and in some studies, a female-advantage has also been found (e.g., Pausewang et al., 2012). Reports of male and female differences in speech rate tend to suggest that males speak faster than females (e.g., Byrd, 1992; 1994; Pepiot, 2014; Whiteside, 1996). However, the overwhelming stereotypical opinion is that females do in fact speak faster than males (e.g., Bond & Feldstein, 1982; Weirich & Simpson, 2014). Hence, manipulations in speech rate may also be important when making judgements about the sex of the speaker. Research suggests that both F0 and speech rate are likely to be important when making estimations about speaker age (e.g., Hartman & Danhauer, 1976; Linville & Fisher, 1985; Ptack & Sander, 1966; Shipp et al., 1992). Nevertheless, at present the literature examining the relationship between perceptual judgements about characteristics of the speaker are still surprisingly sparse.

Furthermore, very few have considered the role of speech rate and whether they are likely to affect perceptual judgements about characteristics of the speaker. Given that speakers are likely to vary the rate at which they speak (refer to Chapter 2, Section 2.3.1 for further detail), further exploration of this cue is necessary. There are also several methodological issues with the studies reviewed, making it difficult for any clear conclusions to be drawn.

Manipulations in F0 and speech rate may also be important for accurate recognition of these cues. Research into the factors that affect recognition performance for the speaker has a long history. Despite this however, remarkably very few have considered the impact of manipulations in F0 or speech rate of the voice and how they can affect recognition performance for these cues. The studies that do exist on this topic have identified that categorical memory processes distort memory for voice F0 and speech rate in a manner where memory is exaggerated (i.e., the accentuation effect (Mullenix et al., 2010; Stern et al., 2007). Whilst insightful, the findings are limited, and there are several methodological issues that make it difficult to generalise the results to other voices in the real-world. For example, research has used only one male voice, and F0 and speech rate manipulations were found to fall considerably outside the typical F0 and speech rate ranges of voiced speech. What is more, there may be other factors that contribute to performance on a memory task for voice F0 and speech rate. These include the time course of echoic memory, and whether the same sentence or a different sentence was spoken to the one previously heard. To date however, no research has explored these ideas further.

3.3.1 Research Questions

The findings of existing studies therefore leave several important questions unanswered. This literature review has highlighted some important gaps in knowledge, which the subsequent experiments seeks to fill.

The specific research questions to be addressed throughout this thesis are as follows:

- **(1):** Do manipulations in fundamental frequency (F0) or speech rate affect perceptual judgments about the paralinguistic characteristics of the speaker, and if so, how do they change? Specifically, how do manipulations in F0 or speech rate affect perceptions of;

  a) speaker identity? (explored further in Experiment 2, Chapter 5)

  b) speaker sex? (explored further in Experiment 3, Chapter 6)

  c) speaker age? (explored further in Experiment 4, Chapter 7)

- **(2):** Do manipulations in fundamental frequency (F0) or speech rate affect recognition performance for voices, and if so, can the findings be explained using the accentuation effect? (explored further in Experiment 5, Chapter 8)

- **(3):** Do listeners become increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition when;

  a) F0 is increased and decreased, and the inter-stimulus interval between presentation of the target voice and the sequential voice pair is increased? (explored in Experiment 6, Chapter 9)

  b) F0 is increased and decreased, and a different sentence is spoken in the sequential voice pair to the one previously spoken by the target voice? (explored in Experiment 7, Chapter 10)

CHAPTER 4. STIMULI DEVELOPMENT

4. Introduction

The following chapter outlines how the voice stimuli were developed for the experiments in this thesis. It begins by explaining how the voices were manipulated and the measurements that were calculated for both F0 and speech rate. It then moves on to discuss several factors that have been found to affect the performance of listeners in speaker perception and recognition tasks, and explains how these have been controlled for throughout the experiments. The chapter also reports several experiments that were carried out to establish the properties of the stimuli and to validate the stimuli used in this thesis.

4.1 Stimuli Development

The voices used in all of the experiments were obtained using Natural Reader 12.0 software (http://www.naturalreaders.com/index.html). Natural Reader is a text-to-speech software application with realistic and natural sounding synthesised voices, generating speech samples from concatenated pieces of real human speech. Synthetic speech was used because of the need for precisely controlled stimuli that varied in F0 or speech rate, and to ensure that all of the voices were unfamiliar to listeners. Six voices were used in total. All voices were English speaking and had a similar southern English accent. This was important as it was necessary to control for regional accent (refer to Section 4.2.2).

The speech samples were created by typing the following phrase *"Spring is the season where flowers appear, summer is the warmest season of the year."*, in Natural Reader. A non-emotive speech phrase was chosen because emotional content has been found to influence memory of a voice (e.g., Hollien, Saletto, & Miller, 1993; Solan & Tiersma, 2003) (refer to Section 4.2.3).

All of the voice samples were manipulated using Audacity® software (http://www.audacityteam.org/). Audacity® is a freely available audio software application that can be used to edit sounds and was chosen to manipulate the voices because it allowed the author to alter one characteristic (e.g., F0) whilst holding the other constant (e.g., speech rate). It was important to control for this to ensure that findings in the experiments were due to manipulation of the characteristic of direct interest. Manipulations to both F0 and speech rate were made using the percentage change tool (refer to Section 4.1.1 and Section 4.1.2 for further detail). Relative, rather than absolute, percentage changes were used to manipulate the voices. Relative percentage change takes into account the overall frequency, or speech rate, of the stimuli being manipulated. Thus, each voice is manipulated by the same percentage in relation to the mean F0. For example, a percentage change of 5% will be smaller for a voice with a mean F0 of 95Hz than for a voice with a mean F0 of 120Hz. This ensured that all manipulations made were proportionally the same across the voices used, and so that findings could be compared with each other. The voice samples were saved as separate .wav files so that they could be used individually in future experiments. The original voices and subsequent manipulations formed the basis of the voice stimuli used in this thesis.

4.1.1 Manipulations in Fundamental Frequency (F0)

Manipulations in F0 were made using the Change Pitch tool in Audacity®. Change Pitch works by applying an up or down percentage change to the existing frequency of a selection. Manipulations were made by increasing and decreasing each voice by 5% and 10%. This resulted in a total of five versions of each voice (i.e., the original version and four manipulated versions). These voices were used in all the experiments (Experiments 1, 2, 3, 4, 5a, 6, and 7).

For Experiments 5a, 6, and 7, the voices were manipulated further to obtain the target and distractor voices required for the experiments. Of the six original synthesised voices, four were used for experimentation (2 male; 2 female). The two male voices and two female voices with the highest naturalness ratings were chosen (this will be discussed further in Section 4.3.3). For each of the original synthesised voices, the 10% manipulated versions were used to obtain target voices in the higher and lower F0 range. All four original voice samples fell within the moderate speaking range for F0, and thus acted as moderate target voices. To obtain the distractor speech samples for Experiments 5a, 6, and 7, each target voice was further increased and decreased by 5%, 7%, and 10%.

For Experiment 7, a different sentence was also used. The speech samples were created by typing the following phrase *"Living cost have more than tripled, and gas has gone down one third."*, in Natural Reader. The same method was used as the one described above to obtain the target and distractor voices speaking this sentence.

Measurements of F0 for the voices were calculated using Praat (Boersma & Weenik, www.praat.org). Praat is a commonly used and freely available software application that can be used to precisely analyse speech sounds. Manipulations in F0 using Praat have also been used by others in their work (Feinberg, Jones, Little, Burt, & Perrett, 2004; Puts, 2005; Puts, Gaulin, & Verdolini, 2006; Wells et al., 2013), and was therefore deemed suitable for the purpose of this thesis. The minimum and maximum F0 parameters were adjusted using the Pitch Range setting. A pitch range between 75 to 600 Hz was deemed appropriate to account for the typical male and female ranges in F0. Measurements were obtained using the Show

Pitch tool. This produces a mean F0 value for the utterance. Figure 4.1 provides an illustration of the mean F0 represented on the spectrogram.

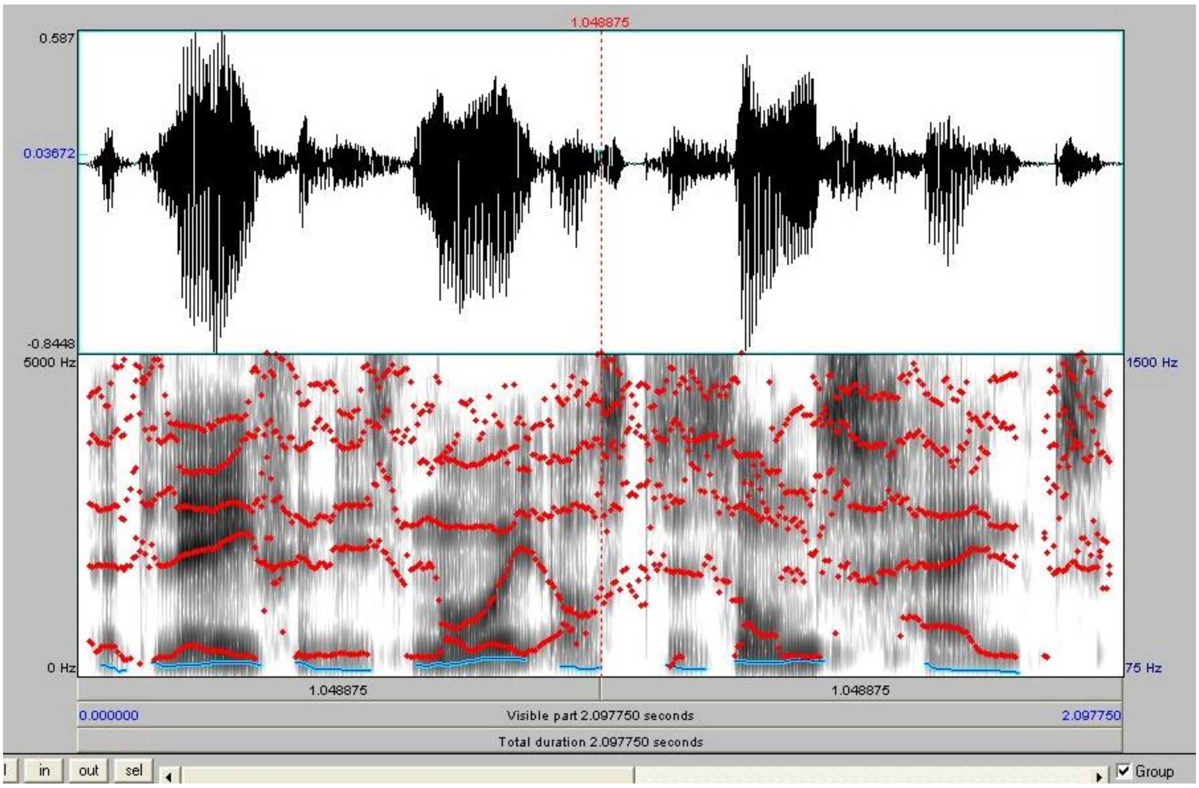


Figure 4.1: Screen shot of a sound wave (upper panel) and spectrogram (lower panel) of the sentence *"Life is beautiful when the sun shines"*, spoken by a male speaker. The degree of blackness is proportional to the amount of energy in that frequency region. The concentration of energy at the lower end of the spectrogram represents the F0 (represented by the pitch contour, i.e., the blue line). Formants are displayed by black bands. The red dotted lines on the spectrogram represent formants.

Table 4.1 presents the mean F0 (in Hz) of each of the original voice samples and their manipulated versions (increased and decreased in F0 by 5% and 10%), listed separately for male and female speakers.

Table 4.1: *Mean Fundamental Frequency (F0; in Hz) of voices listed separately for manipulation (increase or decrease in F0) and sex of speaker (male or female).*

	Male Speakers			Female Speakers		
	Voice One	Voice Two	Voice Three	Voice Four	Voice Five	Voice Six
Manipulation						
+10%	116	123	128	190	228	238
+5%	111	118	122	181	217	228
0% (original)	106	112	116	173	207	217
-5%	100	106	110	165	197	208
-10%	95	101	104	157	186	195

Note: Calculations are shown in Hertz (Hz).

The mean F0 (in Hz) of the target and distractor voices for the sentence *"Spring is the season where flowers appear, summer is the warmest season of the year"* used in Experiments 5a, 6, and 7 are provided in Appendix A1. The mean F0 (in Hz) of the target and distractor voices for the sentence *"Living costs have more than tripled, and gas has gone down one third"* used in Experiment 7 are provided in Appendix A3.

The mean F0 of the different sentences used in Experiment 7 were compared to determine whether they differed in F0. Table 4.2 presents the mean F0 (in Hz) of each of the target voice samples (high: +10%, moderate: 0%, and low F0: -10%), listed separately for male and female speakers, and for the different sentences spoken in Experiment 7.

Table 4.2: *Mean Fundamental Frequency (F0; in Hz) of voices listed separately for each of the target voices (high: +10%, moderate: 0%, or low F0: -10%), sex of voice (male or female), voice (1, 2, 3, or 4), and sentence spoken (one or two).*

	Male Voices				Female Voices			
	Voice 1		Voice 2		Voice 3		Voice 4	
Sentence	One	Two	One	Two	One	Two	One	Two
+10%	116	119	123	125	228	227	238	232
0%	106	108	112	114	207	206	217	211
-10%	95	97	101	103	186	185	195	190

Note: Calculations are shown in Hertz (Hz). Sentence One: *"Spring is the season where flowers appear, summer is the warmest season in the year"*. Sentence Two: *"Living costs have more than tripled, and gas has gone down one third"*. +10% depicts high F0 target voices, 0% depicts moderate F0 target voices, and -10% depicts low F0 target voices.

Table 4.2 shows that whilst the mean F0 of the voice samples do differ for the two sentences used, it was decided upon that this was the most appropriate method of manipulation as it added realistic and real-world variability in the speech samples used (i.e., in a real-world situation, there will be a slight variation in a speakers F0 when a different sentence is spoken).

All manipulations of the voice samples were kept within the typical male and female F0 ranges for voiced speech (i.e., between 80 to 180 Hz for males, and 160 to 255 Hz for females (Baken, 1987; Titze, 1994)).

4.1.1.1 Formant Values

Manipulations in F0 also changed the frequency of the formant values. This was important because changes in F0 made by a speaker in the real world would also effect the frequency of formant values, and the author of the thesis wanted to replicate this situation. Furthermore, research has shown that manipulations in both F0 and formant values affect perceptual judgements of the speaker differently than when manipulations in only one parameter are made (e.g., Coleman, 1971; Whiteside, 1998) (refer to Chapter 3, Section 3.1.2.1.4). Measurements of the formant values were calculated for the voices using Praat. First, the formants were identified on the spectrogram by using the Show Formants tool. The formants are illustrated by the red dotted lines on the spectrogram in Figure 4.1 above. A vowel formant was then selected by clicking on it and dragging the cursor horizontally across the spectrogram until the desired section had been selected. The mid-point of the vowel formant was selected by clicking in the middle of the selection made. Measurements for the formant vowel were then obtained using the Formant Listing tool, which produces a list of the first three formants for the mid-point of the selected vowel. Table 4.3 presents the mean frequency (in Hz) for the first three formants (F1, F2, and F3) for the vowel 'IY' (heard as 'ea') in the word *"season"* from the spoken utterance.

4.1.2 Manipulations in Speech Rate

Manipulations in speech rate were made using the Change Tempo tool in Audacity®. Change Tempo works by applying an up or down percentage change to the existing rate of a selection. Manipulations were made by increasing and decreasing each voice by 5%, 10%, 15%, and 20%. This resulted in a total of nine versions of each voice (i.e., the original version and eight manipulated versions). A greater number of manipulations were made for speech rate than they were for F0 because the same percentage change in F0 lead to a greater perceptual change than it did for speech rate. These voices were used in Experiments 1, 2, 3, 4, and 5b.

For Experiment 5b, the voices were manipulated further in speech rate to obtain the target and distractor voices required for the experiments. Of the six original synthesised voices, four were used for experimentation (2 male; 2 female). The two male voices and two female voices with the highest naturalness ratings were chosen (this will be discussed further in Section 4.3.3). For each of the original synthesised voices, the 20% manipulated versions were used to obtain target voices in the faster and slower in the speech rate range. All four original voice samples fell within the moderate speaking range for speech rate, and thus acted as moderate target voices. To obtain the distractor speech samples for Experiments 5, each target voice was further increased and decreased by 10%, 12%, and 20%.

Measurements of speech rate are influenced by the inclusion and exclusion of pauses and hesitations. Laver (1994) distinguishes between speech rate and articulation rate. Speech rate refers to the rate of speech for the whole speaking-turn and includes all speech material together with any silent pauses (Laver, 1994). Measurement of articulation rate includes all audible speech material but excludes silent pauses. Whilst excluding pause time more closely conveys the pace at which speech is produced, it does not take into account speaker-specific ways of transmitting information such as pauses and hesitations (Jacewicz, Foz & Wei, 2010). For this thesis, speech rate included pause time so that any speaker-specific ways of transmitting information were incorporated in the voices used. The inclusion of pauses is also likely to reflect differences that exist between voices that are heard in the real world. Speech rate will be defined as the number of output units (in syllables) per unit of time, including pause intervals that may separate uninterrupted articulatory sequences (Crystal & House, 1986, 1988; Miller, Grosjean, & Lomanto, 1984).

Table 4.3: *Mean frequency (in Hz) for the first three formants (F1, F2, and F3) for the vowel 'IY' (heard as 'ea') in the word "season" from the spoken utterance.*

	Formant Value	-10%	-5%	0% (original)	5%	10%	Average Vowel Formant
Voice 1 (male)	F1	198	223	234	248	262	*270*
	F2	2066	2141	2267	2358	2455	*2300*
	F3	2796	2767	2820	2807	2776	*3000*
Voice 2 (male)	F1	238	242	257	269	283	*270*
	F2	2224	2370	2454	2593	2709	*2300*
	F3	2686	2769	2778	2819	2987	*3000*
Voice 3 (male)	F1	267	278	293	**309**	**318**	*270*
	F2	2074	2185	2281	2350	2462	*2300*
	F3	2510	2667	2718	2779	2876	*3000*
Voice 4 (female)	F1	**256**	271*	357	388	414	*300*
	F2	2369*	2463*	2673	2793	2932	*2800*
	F3	**2444**	**2657**	3121	3239	3403	*3300*

Voice 5 (female)	F1	276*	306	311	325	343	*300*
	F2	**2203**	2614	2160	2345	2598	*2800*
	F3	**2681**	2657	2736	2918	2989	*3300*
Voice 6 (female)	F1	295	315	337	360	376	*300*
	F2	**2294**	2445	2550	2639	2749	*2800*
	F3	**2483**	2758	2952	3028	3143	*3300*

Note: Calculations in bold depict formant values that fall outside the typical F0 range for male or female voiced speech (depending on whether the voice is male or female). Calculations with an asterisk (*) depict formant values close to the typical formant average of the opposite sex.

Various units of measurement have also been considered as a basis for measurement of speech rate. One common method is words per minute (wpm). However, calculations using syllables rather than words are often considered as being a more accurate and reliable estimate of the rate of speech (Dlugen, 2012). This is because calculations using words are dependent upon the length of the words spoken in the spoken sentence, and not all words in the English language are equal. For example, consider the following two sentences (taken from http://sixminutes.dlugan.com/speaking-rate/):

1. *'Modern readability tests are designed to indicate comprehension difficulty when reading a passage of contemporary academic English'.* (17 words; 41 syllables).

2. *'Ask not what your country can do for you; ask what you can do for your country'.* (17 words; 19 syllables).

If a person were to speak these two sentences at the same rate in words per minute, the first sentence using longer words would seem considerably faster than the second sentence using shorter words because more is being spoken (Dlugen, 2012). It was therefore decided upon to use syllables per second (syll/sec) for all calculations of speech rate.

All measurements of speech rate were calculated by hand using the following formula,

$$\text{Speech Rate (syll/sec)} = \frac{\text{Total number of syllables in utterance}}{\text{Number of seconds of utterance}}$$

where *total number of syllables in utterance* refers to the number of perceptually fluent syllables in the utterance (Chon, Ko, & Shin, 2004), and *number of seconds of utterance* refers to the length of the chosen sentence in seconds, including all pauses. The length of the utterance was sourced from Audacity®.

Table 4.4 presents the speech rate (in syll/sec) of each of the original speech samples and their manipulated versions (increased and decreased in speech rate by 5%, 10%, 15%, and 20%).

Table 4.4: *Speech rate (in syll/sec) of each of the original speech samples and their manipulated versions (i.e., increased and decreased in speech rate by 5%, 10%, 15%, and 20%).*

	Male Speakers			Female Speakers		
	Voice One	Voice Two	Voice Three	Voice Four	Voice Five	Voice Six
Manipulation						
+20%	3.94	4.31	4.31	2.54	4.29	4.33
+15%	3.78	3.75	4.12	2.69	4.11	4.14
10%	3.62	3.59	3.95	2.85	3.93	3.96
+5%	3.45	3.42	3.77	3.01	3.75	3.79
0% (original)	3.29	3.26	3.59	3.17	3.58	3.62
-5%	3.12	3.10	3.41	3.31	3.40	3.42
-10%	2.96	2.93	3.23	3.49	3.22	3.24
-15%	2.79	2.77	3.05	3.65	3.04	3.06
-20%	2.63	2.60	2.87	3.80	2.86	2.88

Note: Calculations are shown in syllables per second (sps).

The mean speech rate (in syll/sec) of the target and distractor voices for the sentence *“Spring is the season where flowers appear, summer is the warmest season of the year”* used in Experiment 5b are provided in Appendix A2.

The manipulations were kept very close to the typical male and female speech rate ranges (i.e., between 3.3 to 5.9 syll/sec (Arnfield, Roach, Setter, Greasley, & Horton, 1995: Tsao & Weismer, 1997)).

4.2 Controlling for Extraneous Variables in the Stimuli

The following section outlines several factors that have been found to affect the performance of listeners in speaker perception and recognition tasks, and explains how these have been controlled for during the experiments.

4.2.1 Speaker Familiarity

The voices used in this thesis were unfamiliar to the listeners. It is important to make the distinction between familiar and unfamiliar voices because the ability to recognise familiar speakers from their voice alone has been found to be superior. Research has shown that the processes for identifying familiar and unfamiliar speakers are distinctly different from each other and are located in different regions of the brain (e.g., Schmidt-Nielsen & Stern, 1985; Van Lacker, Kreiman, & Emmorey, 1985; Van Lacker, Kreiman, & Wickens, 1985; Yarmey, Yarmey, & Yarmey, 2001). Several studies have shown that familiar listeners perform significantly better than naïve listeners (i.e., listeners who do not know the speakers), when identifying the same speakers (e.g., Amino & Arai, 2009; Foulkes & Barron, 2000; Rose & Duncan, 1995; Wenndt, 2016; Yarmey, Yarmey, & Yarmey, 2001), with identification rates of familiar voices found to be as high as 97% to 99% (Hollien, Majewski, & Doherty, 1982; LaRiviere, 1972). Listeners have also been found to be better at recognising familiar speakers than unfamiliar speakers using only one word, *"hello"* (Ladefoged & Ladefoged, 1980).

4.2.2 Ethnicity, Other Race, and Accented Voices

The voices used in this thesis were standardised to ensure that they were all English speaking and had a similar regional accent. This was important because studies on own- and other-race/ethnicity in voices have shown that people may be better at recognising voices of their own race/ethnicity than those of another race/ethnicity. For example, research has found that the identification of a speaker is significantly improved when listeners are familiar with the language being spoken, in contrast to when statements are spoken in a foreign language (Doty, 1998; Goggin, Thompson, Strube, & Simental, 1991; Hollien, Majewski, & Doherty, 1982; Koster & Schiller, 1997; Koster, Schiller, Kunzel, 1995; Schiller & Koster, 1996).

Similarly, the other-accent effect suggests that listeners are better able to recognise speakers with a more familiar (or similar) accent. The effect of trying to recognise voices speaking with an accent has been investigated in several studies. Research has shown that it is more difficult to recognise a speaker with an accent than one without an accent. For example, Goldstein, Knight, Bailis, and Conover (1981) showed that voices speaking English with a Chinese or Black American accent were not as well recognised as voices with a general American English accent. Furthermore, Thompson (1987) demonstrated that Spanish-accented English voices were recognised more poorly than English accented voices. Australian listeners have also been shown to have an impairment when recognising speakers with an unfamiliar (British English) accent than when recognising a speaker with a familiar (Australian English) accent (Vanags, Carrol, & Perfect, 2005).

4.2.3 Emotional Stress and Arousal

The sentences used in this thesis were non-emotive and spoken in a normal, conversational tone. This is important because voices convey information about the emotional state of the speaker. However, the effects of emotionality/stress on memory have received little attention from researchers. If a speaker is experiencing stress, anger, or anxiety, this will be reflected in various speech characteristics, such as F0, speech rate, duration, and number of speech bursts (Hollien, Saletto, & Miller, 1993). Accurate identification of a speaker is poorer if they use a different tone to the one spoken at the encoding stage. For example, Solan and Tiersma (2003) report a case in which a rapist was very calm and soft-spoken while committing the assault. Later, when the victim was asked to identify the voice, they failed to make a positive identification when the suspect was speaking in an angry and abusive tone. Nevertheless, when the suspect spoke in a calm voice, the victim claimed to recognise the voice immediately.

4.2.4 Voice Sample Durations

The duration for all sentences spoken during the experiments were the same in length. Research has shown that when listeners are given a longer opportunity to hear someone speak, they are more likely to accurately identify the speaker than when they are given less time to hear someone speak (Cook & Wilding, 1997; Hammersley & Read, 1985; Orchard & Yarmey, 1995; Read & Craik, 1995; Yarmey, 1991; Yarmey & Matthys, 1992). This is likely because listeners have longer to attend to the voice and make decisions about it (Roebuck & Wilding, 1993). Thus, it was important to ensure that the duration of the sentences spoken during the experiments were the same so that listeners were exposed to the voices for the same amount of time, and had the same amount of time to make any decisions.

4.3 Stimuli Validation

This section reports five experiments that were carried out to obtain information about the voices, and to ensure that the stimuli used for the experiments were appropriate.

4.3.1 Experiment 1a: Perceived Similarity in Fundamental Frequency (F0) and Speech Rate of the Voice Stimuli

4.3.1.1 Introduction and Aims

Experiments 2, 5, 6, and 7 in this thesis involved the presentation of voices in succession in a sequential voice pair. Thus, it was important to determine the transition threshold for changes in either F0 or speech rate for the voice samples. Experiment 1a set out to investigate at what point manipulations in either F0 or speech rate of the original voice samples were perceived as sounding different from the original (i.e., unmanipulated) voices.

It should be noted that whilst manipulations in F0 also changed the frequency of the formant values, it was deemed uneccessary to determine the transition threshold for changes in formant frequencies for the voice samples. The aim of this thesis was to investigate the effect of manipulations in F0 on perceptions about the speaker and recognition performance for the voice. As previously acknowledged, in the real world, changes made by a speaker in F0 also effects the frequency of formant values, and the author of the thesis wanted to replicate this situation. Research has considered the role of F0 and formant frequencies separately, where only one of these properties is manipulated to determine their effect on perceptions about the speaker (e.g., Coleman, 1971; Gelfer & Mikos, 2005; Hilenbrand & Clark, 2009). If this had been the case, it would have also been necessary to establish at what point listeners are able to detect changes in the frequencies of the formant values. Nevertheless, because this methodology was not employed in this thesis, it was deemed unnecessary to undergo any formal testing for this.

For the purposes of the research, it was decided upon that the most appropriate methodology to use would be the method of constant stimuli. This approach is least sensitive to response bias as it leaves the subject uncertain about the size of the signal to be presented next and a more realistic sensory threshold can be obtained. The method of constant stimuli requires a fixed set of stimuli to be developed beforehand. The levels of the stimuli are not related from one trial to the next, but are instead presented in a random, or semi-random, order (Green & Swets, 1988). The listeners are presented with a constant comparison stimulus (i.e., the unmanipulated versions) and one of the varied stimuli, and asked to determine whether the voices sound the same or different. The data obtained can be plotted as a psychometric function where the proportion of times the signal is detected (i.e., the voices sound different from each other) is plotted as a function of signal magnitude. The stimulus difference that is noticed and elicits a positive response (i.e., voices sound different from each other) at some fixed percentage of the time can then be calculated. The value of the signal corresponding to 50% response (the sensory threshold) is most typically used (Green & Swets, 1988).

4.3.1.2 Method

4.3.1.2.1 Participants

A total of 72 undergraduate students (36 males; 36 females) were recruited from Nottingham Trent University and received course credit for their participation. The ages of the participants ranged from 19 to 27 years old (*M* = 22.71 years, *SD* = 3.70 years). The inclusion criteria for the study required individuals to be between 18-30 years of age, had no known hearing deficits, had English as their first language, had not undergone any musical training[1], and had not heard the stimuli used in the experiment before.

4.3.1.2.2 Materials and Stimuli

All six of the original voices and their subsequent manipulations (i.e., *For F0:* increased/decreased by 5% and 10%; *For Speech Rate:* increased/decreased by 5%, 10%, 15%, and 20%) were used for the experiment. The speech samples were presented binaurally using Sony dynamic stereo headphones (Model No. MDR-V150). The experiment was run on a Sony Vaio laptop computer (Model No. SVF153B1YM) using PsychoPy version 1.7701 (Peirce, 2007) to control the presentation of the voices and collect participant responses.

4.3.1.2.3 Procedure

The experiment consisted of two parts; the F0 condition and the speech rate condition (counterbalanced across participants). As illustrated in Figure 4.2, using a perceptual discrimination paradigm, the participants were given a 2AFC (same/different key press) voice

[1] Musicians with extensive musical training have been found to outperform non-musicians on speech perception and unfamiliar voice identification tasks (e.g., Bregman & Creel, 2014; Parbery-Clark, Skoe, Lam, & Kraus, 2009; Slec & Miyake, 2006). Thus, it was important to ensure that no participants had undergone any musical training as this may have impacted upon the findings.

discrimination task. In each trial, the original target voice was used as the standard stimulus and presented on all trials. The standard stimulus was paired with either itself or a manipulated version (increased/decreased in F0, or increased/decreased in speech rate) and presented in a random order. The stimuli were presented as within voice pairs with a 1 second inter-stimulus interval between each voice (e.g., original voice four and increased by 5% voice four). Following presentation of each trial, the participants had to indicate whether the two voices sounded the same or different by pressing either the left or right keys on the laptop keyboard. For the F0 condition, there were 60 trials in total (5 trials for each voice, with each trial being presented twice). For the speech rate condition, there were 108 trials in total (9 trials for each voice, with each trial being presented twice). The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life. Upon completion of the experiment, the participants were fully debriefed and thanked for their time and participation.

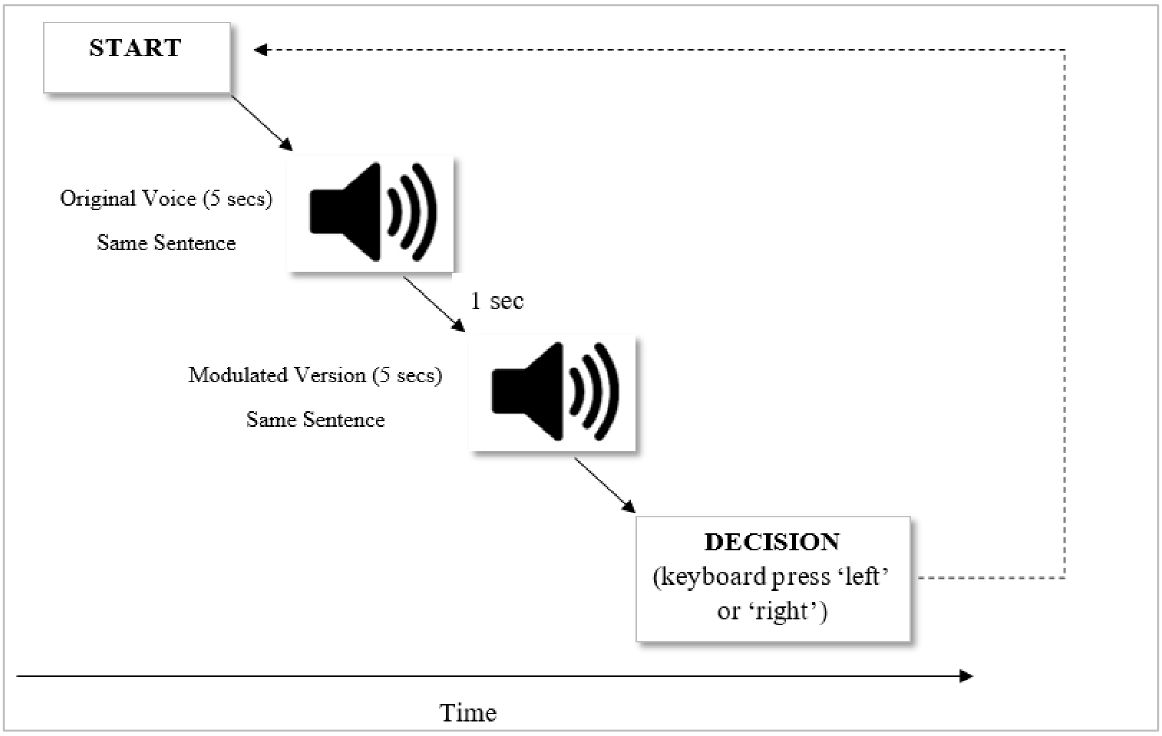


Figure 4.2: An illustration of the procedure in Experiment 1a.

4.3.1.3 Results and Discussion

Same/different performance was susceptible to relatively smaller changes in F0 than for speech rate (i.e., there appears to be a greater tolerance for changes in speech rate than for F0). Specifically, a change in speech rate almost double that of a change in F0 is required before listeners are able to detect the voice as sounding different from the original version. This is supported by others who have found greater changes in speech rate than F0 are required to influence similarity ratings of the speaker (Gelfer, 1993; Murry & Singh, 1980; Singh & Murry, 1978).

4.3.1.3.1 Fundamental Frequency (F0)

Figure 4.3 depicts the mean percentage of times listeners heard the manipulated versions of the voices as sounding the same as the original (i.e., unmanipulated) versions. The data were collapsed across plus and minus manipulations owing to the comparative effect an increase and decrease in manipulation had on same/different ratings (refer to Appendix B1). The results suggested that greater manipulations in F0 increased the likelihood that the voices sounded different to the original version of the voice. The transition threshold (i.e., the value of the signal corresponding to 50% response, or chance level) for all six voices was 6.3%, indicating that voices manipulated by 6.63% or more in F0 were perceived as sounding different from the original version of the voice.

4.3.1.3.2 Speech Rate

Figure 4.4 depicts the mean percentage of times listeners heard the manipulated versions of the voices as sounding the same as the original (i.e., unmanipulated) versions. Data was collapsed across plus and minus manipulations due to the comparative effect an increase and decrease in manipulation had on same/different ratings (refer to Appendix B2). The results suggested that greater manipulations in speech rate increased the likelihood that the voices sounded different to the original version of the voice. The transition threshold (i.e., the value of the signal corresponding to 50% response, or chance level) for all six voices was 11.42%, indicating that voices manipulated by 11.42% or more in speech rate will be perceived as sounding different from the original version of the voice.

It should be noted that whilst a greater percentage change is required in speech rate than it is in F0 for the voices to sound different to the original version of the voice, performance is similar across manipulations in F0 and speech rate, and thus supporting the reasoning behind greater manipulations being made for speech rate than for F0.

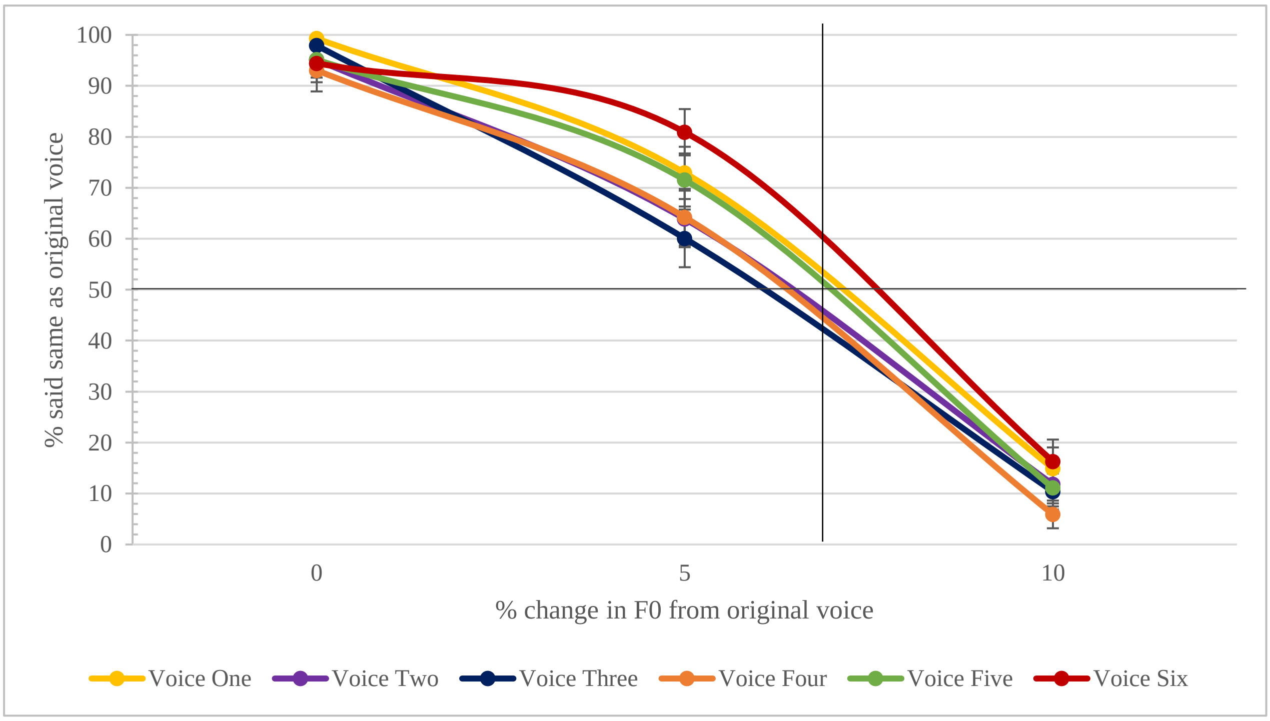


Figure 4.3: Line graph depicting percentage of times listeners heard the voice as the same as the original voice, for F0. Average PSE (taken at 50%, chance level) is 6.63%. Each of the six voices are depicted by a different line colour. 95% confidence intervals are also shown.

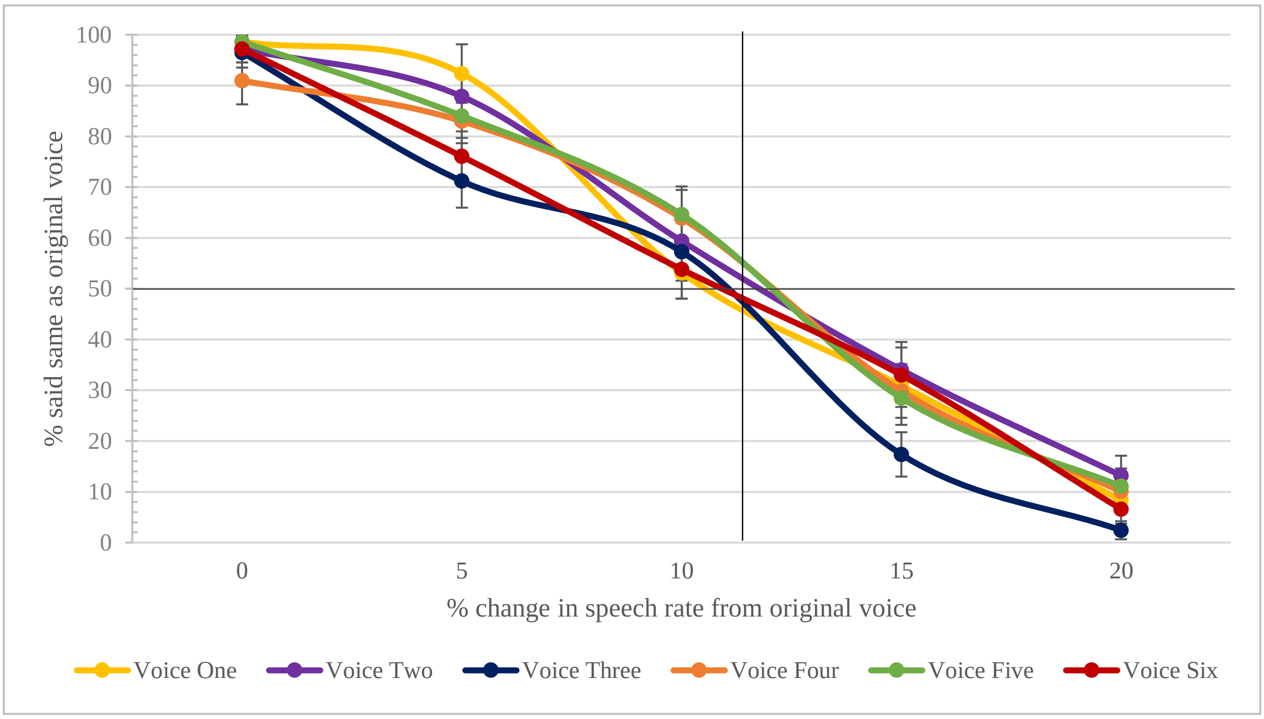


Figure 4.4: Line graph depicting mean percentage of times listeners heard the voice as the same as the original voice, for speech rate. Average PSE is 11.42%. Each of the six voices are depicted by a different line colour. 95% confidence intervals are also shown.

4.3.2 Experiment 1b: Between Speaker Identity Discrimination

4.3.2.1 Introduction and Aims

It was important to ensure that all voices used in this thesis were distinct from each other and were perceived as being different identities so that they would not be confused with another voice that had previously been heard. Therefore, the aim of Experiment 1b was to determine whether the six voices used in Experiments 2, 3, 4, 5, 6, and 7 were perceived as being different speakers (i.e., different identities).

4.3.2.2 Method

4.3.2.2.1 Participants

A total of 72 undergraduate students (36 males; 36 females) were recruited from Nottingham Trent University and received course credit for their participation. The ages of the participants ranged from 18 to 27 years old (*M* = 21.53 years, *SD* = 3.92 years). The inclusion criteria for the study required individuals to be between 18-30 years of age, had no known hearing deficits, had English as their first language, had not undergone any musical training, and had not heard the stimuli used before.

4.3.2.2.2 Materials and Stimuli

All six of the original voice samples (three male, three female) were used for the experiment.

4.3.2.2.3 Procedure

Using a perceptual discrimination paradigm, the participants were given a 2AFC (same/different key press) voice discrimination task. As illustrated in Figure 4.5, in each trial, the stimuli were presented as between voice pairs, whereby an original voice was paired with a different original voice (e.g., original voice 1 and original voice 3) and presented in a random order. There was a 1 second inter-stimulus interval between presentation of each voice. Following presentation of each trial, the participants had to indicate whether the two voices were the same person talking or a different person talking by pressing either the left or right keys on the laptop keyboard. There were 30 trials in total (each original voice being paired with all other original voices, with each trial being presented twice). The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life. Upon completion of the experiment, participants were fully debriefed and thanked for their time and participation.

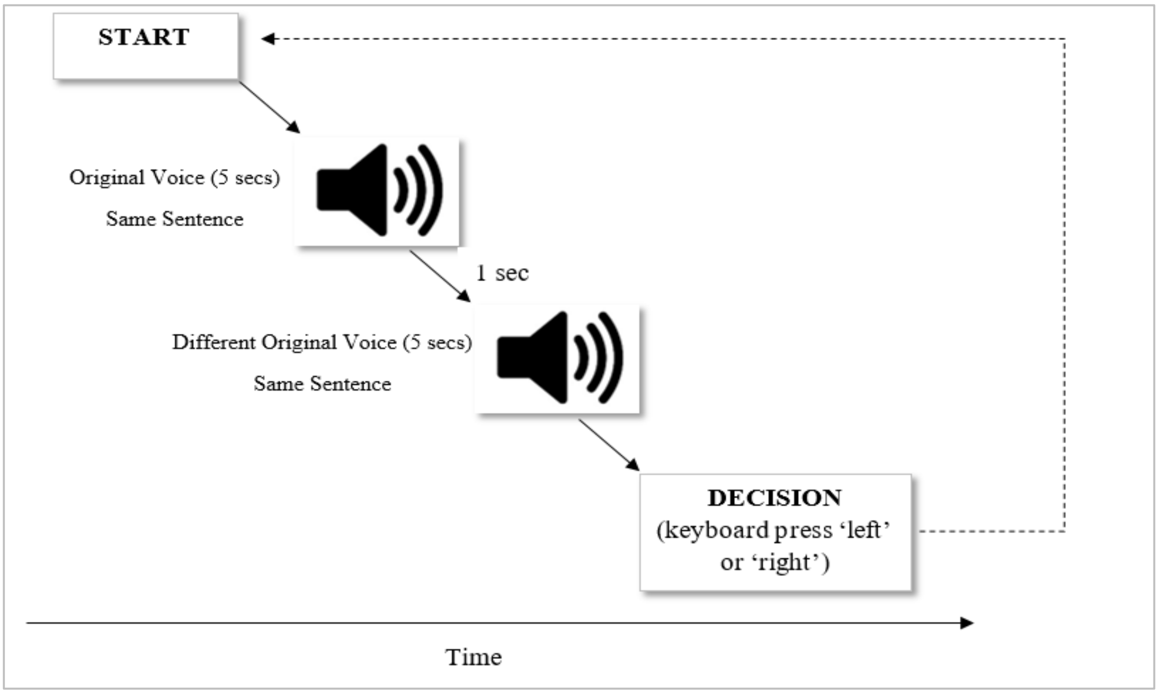


Figure 4.5: An illustration of the procedure in Experiment 1b.

4.3.2.3 Results and Discussion

Table 4.5 presents the mean percentage of time that listeners heard an original version of a voice as sounding like a different speaker compared to another original version of a voice.

Table 4.5: *Mean percentage of times listeners heard an original voice paired with a different original voice as a different speaker.*

	Male Speakers			Female Speakers		
	Voice One	Voice Two	Voice Three	Voice Four	Voice Five	Voice Six
Voice One	0.69	98.61	98.61	100	98.61	99.31
Voice Two	98.61	0	97.22	97.91	98.61	97.92
Voice Three	98.61	97.22	0	100	100	100
Voice Four	100	98.61	100	0.69	98.61	99.31
Voice Five	98.61	98.61	100	98.61	0.69	100
Voice Six	99.31	97.92	100	99.31	100	0

Note: Calculations are shown as a percentage (%).

The results showed that listeners could correctly determine that the voices were different speakers with almost 100% accuracy. The listeners were also able to correctly determine when the voices were the same speaker with almost 100% accuracy. Therefore, we can assume that the voices used for the experiments in this thesis are distinct from each other and perceived as being different speakers.

4.3.3 Experiment 1c: Naturalistic Ratings of Voices used in Experiments 2, 3, and 4

4.3.3.1 Introduction and Aims

It was important to ensure that all voices used in this thesis were generalizable to those voices that are heard in a real-world environment. Therefore, the aim of Experiment 1c was to determine the extent to which the synthesised voices used in Experiment's 2, 3, and 4 sounded like real voices.

4.3.3.2 Method

4.3.3.2.1 Participants

A total of 20 undergraduate students (10 males; 10 females) were recruited from Nottingham Trent University and received course credit for their participation. The ages of the participants ranged from 18 to 29 years old (*M* = 20.34 years, *SD* = 3.67 years). The inclusion criteria for the study required individuals to be between 18-30 years of age, had no known hearing deficits, had English as their first language, had not undergone any musical training, and had not heard the stimuli used before.

4.3.3.2.2 Materials and Stimuli

The stimuli and materials were identical to that used in Experiment 1a. All six of the original voices and their subsequent manipulations (i.e., *for F0:* increased and decreased by

5% and 10%; *for speech rate:* increased and decreased by 5%, 10%, 15%, and 20%) were used for the experiment.

4.3.3.2.3 Procedure

As illustrated in Figure 4.6, in each trial, the participants were presented with one of the six original voices or manipulated versions of the voices (*for F0:* +/- 5%, +/- 10% and *for speech rate:* +/-5%, +/-10%, +/-15%, +/-20%), uttering the same speech phrase. Each voice was presented one at a time, and in a random order. There were 114 trials in total (with each voice being presented twice). After presenting each voice, the participant's task was to decide how natural sounding the voices were from 1 to 10 (with '1' being not at all realistic and natural sounding, and '10' being very realistic and natural sounding). The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life.

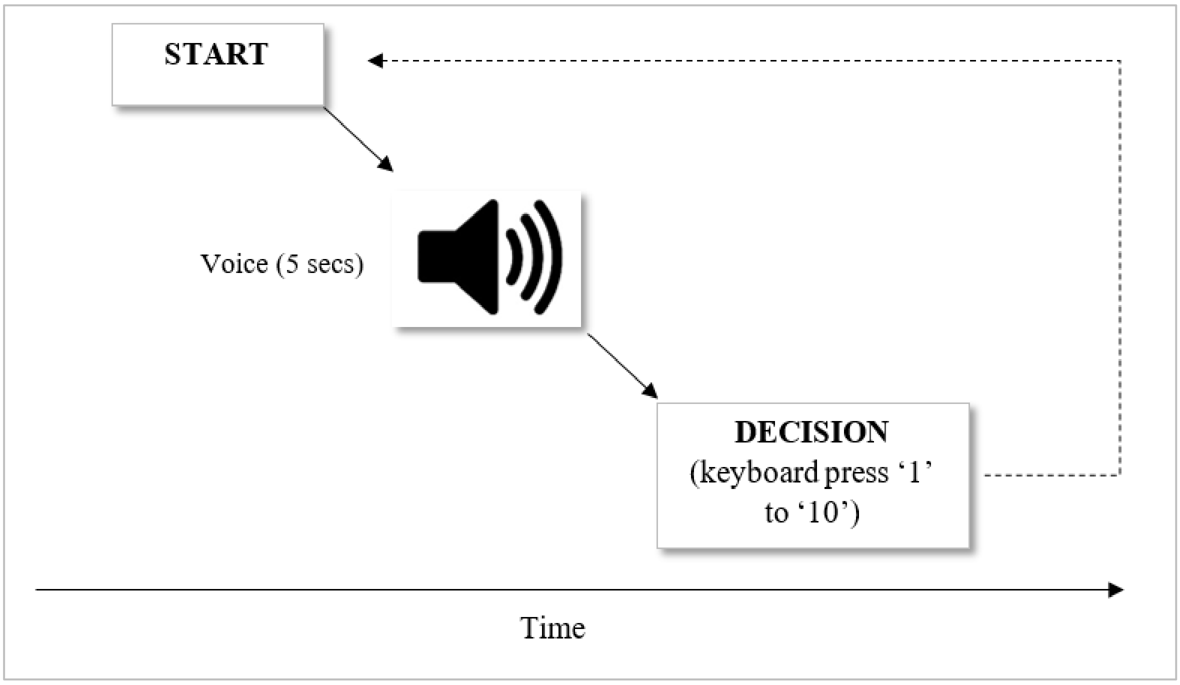


Figure 4.6: An illustration of the procedure in Experiment 1c.

4.3.3.3 Results and Discussion

The mean naturalness ratings were calculated for each voice to determine how realistic, or lifelike, the voices sounded to listeners.

4.3.3.3.1 Fundamental Frequency (F0)

Table 4.6 presents the mean naturalness ratings of the manipulated and unmanipulated (i.e., original) versions of the voices for F0.

Table 4.6: *Mean naturalness ratings of voices listed separately for manipulation (increase or decrease in F0) and sex of speaker (male or female).*

	Male Speakers			Female Speakers		
	Voice One	Voice Two	Voice Three	Voice Four	Voice Five	Voice Six
Manipulation						
+10%	70.25	70	70.25	74.75	73	75.25
+5%	73.75	72.75	73.75	70.75	75.25	73.75
0% (original)	77	73.25	73	74.75	76.5	78.25
-5%	73	75.50	72.75	68	70.75	73
-10%	73.75	73.75	70.50	73.50	72.75	72.25
	73.5	**73.05**	**72.05**	**72.35**	**73.65**	**74.5**

Note: Calculations are shown as a percentage (%).

4.3.3.3.2 Speech Rate

Table 4.7 presents the mean naturalness ratings of the manipulated and unmanipulated (i.e., original) versions of the voices for speech rate.

Table 4.7: Mean naturalness ratings of voices listed separately for manipulation (increase or decrease in speech rate) and sex of speaker (male or female).

	Male Speakers			Female Speakers		
	Voice One	Voice Two	Voice Three	Voice Four	Voice Five	Voice Six
Manipulation						
+20%	75.25	77	73	73	68.75	74
+15%	72.25	74.25	70	73	71.75	75.75
10%	73.75	74.50	72.50	72.50	74.25	75
+5%	74.5	72.25	71.75	75.75	72.25	76
0% (original)	77	73.25	73	76.50	74.75	77
-5%	71.75	73.50	70.25	70.75	71.25	78.25
-10%	68.75	73.50	71.25	70.50	74	69
-15%	73	74.25	70.50	72.25	72	72.25
-20%	71	73.25	68.75	71.50	69	70
	73.03	**73.97**	**71.22**	**72.86**	**72.00**	**73.75**

Note: Calculations are shown as a percentage (%).

The results showed that mean naturalness ratings averaged above 70% for almost all of the synthesised voices. This was observed for the original and manipulated versions of the voices. These values are similar to those identified in the literature (e.g., Jreige, Patel, & Bunnell, 2009) and are a good indication that the synthesised voices used for the experiments were representative of real voices. It should also be noted that the voice samples contained smooth formant transitions and there were no intonational irregularities or prosodic mismatches across words.

4.3.4 Experiment 1d: Naturalness Ratings of Voices used in Experiments 5a, 5b, 6, and 7

4.3.4.1 Introduction and Aims

The aim of Experiment 1d was to determine the extent to which the synthesised voices used in Experiment 5a, 5b, 6, and 7 sounded like real voices. This is important because the author wanted to ensure that the voices used were generalizable to those voices that are heard in a real-world environment.

4.3.4.2 Method

4.3.4.2.1 Participants

A total of 20 participants (10 males; 10 females) were recruited from Nottingham Trent University and received course credit for their participation. The ages of the participants ranged from 18 to 29 years old (*M* = 20.34 years, *SD* = 3.67 years). The inclusion criteria for the study required individuals to be between 18-30 years of age, had no known hearing deficits, had English as their first language, had not undergone any musical training, and had not heard the stimuli used in the experiment before.

4.3.4.2.2 Materials and Stimuli

The target voices (*for F0:* high, moderate, or low F0, and *for speech rate:* fast, moderate, or slow speech rate) and the distractor voices (*for F0:* +/- 5% and +/- 10%; *for speech rate:* +/-10% and +/-20%) for four of the six original voices were used for the experiment.

4.3.4.2.3 Procedure

As illustrated in Figure 4.7, in each trial, the participants were presented with one of the target voices (*for F0:* high, moderate, or low F0; *for speech rate:* fast, moderate, or slow speech rate) or the distractor voices (*for F0:* +/- 5%, +/- 10%; *for speech rate:* +/-10%, +/-20%), uttering the same speech phrase. Each voice was presented one at a time, and in a random order. There were 60 trials in total (6 different target voices, each with 4 distractor voices, with each voice being presented twice). After presenting each voice, the participant's task was to decide how natural sounding the voices were from 1 to 10 (with '1' being not at all realistic or natural sounding, and '10' being very realistic and natural sounding). The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life.

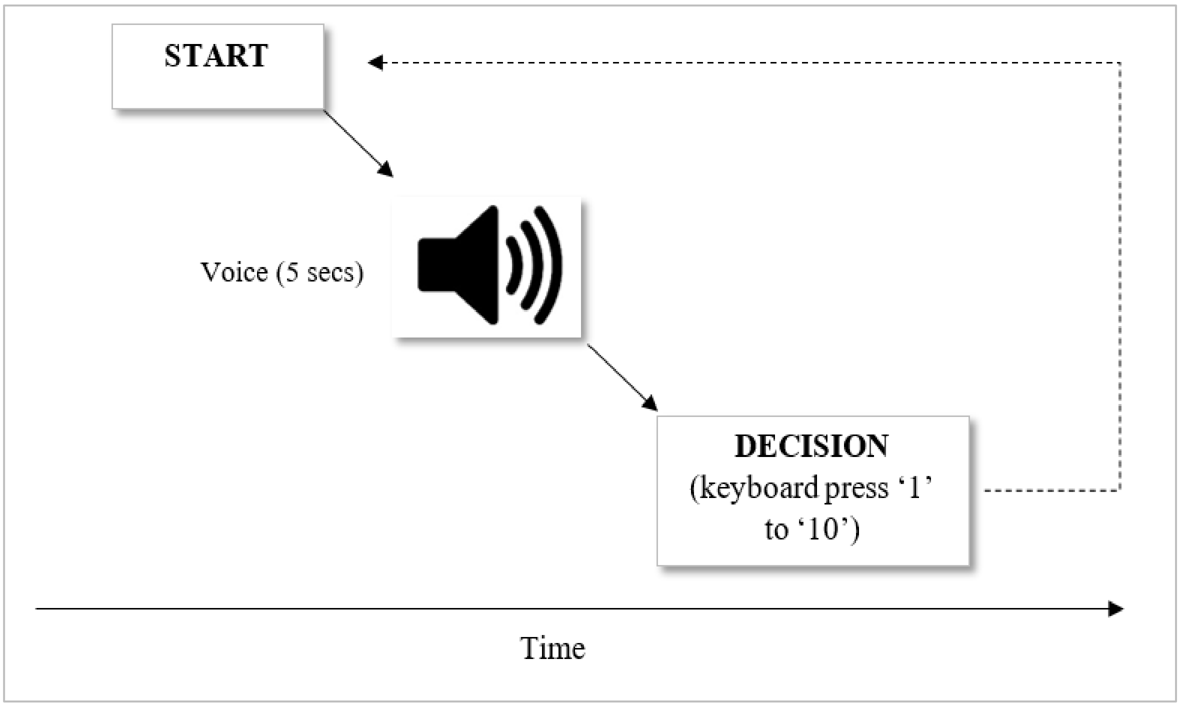


Figure 4.7: An illustration of the procedure in Experiment 1d.

4.3.4.3 Results and Discussion

The mean naturalness ratings were calculated for each voice to determine how realistic, or lifelike, the voices sounded to listeners.

4.3.4.3.1 Fundamental Frequency (F0)

Table 4.8 presents the mean naturalness ratings of the target and distractor voices for F0.

Table 4.8: *Mean naturalness ratings of voices listed separately for target voice, manipulation (increase or decrease in F0) and sex of speaker (male or female).*

	Male Speakers		Female Speakers	
	Voice One	Voice Two	Voice Three	Voice Four
Manipulation				
+10%	75	76	71.75	77
+5%	73.25	75	74.25	75.25
High F0 Target Voice	70	70	73	75.25
-5%	72.25	69.75	75.50	75
-10%	72.50	71.75	74.75	76.50
	72.60	**72.50**	**73.85**	**76.50**
+10%	70.25	70	73	75.25
+5%	73.75	72.75	75.25	73.75
Moderate F0 Target Voice	77	73.25	76.50	78.25
-5%	73	75.50	70.75	73
-10%	73.75	73.75	72.75	72.25
	73.50	**73.05**	**73.65**	**74.50**
+10%	74	72.50	77.75	72
+5%	76	73.25	72.25	73
Low F0 Target Voice	73.75	73.75	72.25	72.25
-5%	72.50	72.75	72	68.50
-10%	72.25	74.25	73	71
	73.70	**73.30**	**73**	**71.35**

Note: Calculations are shown as a percentage (%).

The results showed that mean naturalness ratings averaged at 73.46% for all of the synthesised voices for F0. These values are similar to those identified in the literature (e.g., Jreige et al., 2009) and are a good indication that the synthesised voices used for the experiments were representative of real voices. The voice samples contained smooth formant transitions and there were no intonational irregularities or prosodic mismatches across words.

4.3.4.3.2 Speech Rate

Table 4.9 presents the mean naturalness ratings for the target voices and the distractor voices for speech rate.

Table 4.9: *Mean naturalness ratings of voices listed separately for target voice, manipulation (increase or decrease in speech rate) and sex of speaker (male or female).*

	Male Speakers		Female Speakers	
	Voice One	Voice Two	Voice Three	Voice Four
Manipulation				
+20%	73.25	70.25	71.50	72.75
+10%	73.75	69	72.25	71.50
Fast Rate Target Voice	75.25	77	68.75	74
-10%	72.25	72	72.25	70.25
-20%	74	70.25	68.50	68.75
	73.70	**71.70**	**70.65**	**71.45**
+20%	75.25	77	68.75	74
+10%	73.75	74.50	74.25	75
Moderate Rate Target Voice	77	73.25	74.75	77
-10%	68.75	73.50	74	69

-20%	71	73.25	69	70
	73.03	**73.97**	**72.00**	**73.75**
+20%	70.50	70.50	70.50	73
+10%	74.25	73.50	90	73.75
Slow Rate Target Voice	71	73.25	69	70
-10%	71.50	72.75	72	68.50
-20%	73.75	72.75	69.25	72
	72.20	**72.55**	**74.15**	**71.45**

Note: Calculations are shown as a percentage (%). Calculations in bold depict overall mean naturalness rating for the voices.

The results showed that mean naturalness ratings averaged at 72.60% for all of the synthesised voices for speech rate. These values are similar to those identified in the literature (e.g., Jreige et al., 2009) and are a good indication that the synthesised voices used for the experiments were representative of real voices. The voice samples contained smooth formant transitions and there were no intonational irregularities or prosodic mismatches across words.

4.3.5 Experiment 1e: Validation of the Target Voice Stimuli used in Experiments 5a, 5b, 6, and 7

4.3.5.1 Introduction and Aims

Each of the target voices samples for F0 (i.e., high, moderate, and low F0) and speech rate (i.e., fast, moderate, and slow rate) were tested to determine whether the voices were perceived as being either high, moderate, or low in F0, or fast, moderate, or slow in speech rate.

4.3.5.2 Method

4.3.5.2.1 Participants

A total of 20 participants (10 males; 10 females) were recruited from Nottingham Trent University and received course credit for their participation. The ages of the participants ranged from 18 to 29 years old (*M* = 22.78 years, *SD* = 2.01 years). The inclusion criteria for the study required individuals to be between 18-30 years of age, had no known hearing deficits, had English as their first language, had not undergone any musical training, and had not heard the stimuli used in the experiment before.

4.3.5.2.2 Materials and Stimuli

The target voices (*for F0:* high, moderate, or low F0, and *for speech rate:* fast, moderate, or slow speech rate) for four of the six original voices were used for the experiment.

4.3.5.2.3 Procedure

As illustrated in Figure 4.8, in each trial, the participants were presented with one of the target voices (*for F0:* high, moderate, or low F0, and *for speech rate:* fast, moderate, or slow speech rate) uttering the same speech phrase. Each voice was presented one at a time, and in a random order. There were 48 trials in total (4 different target voices, each with 6 target voices, with each voice being presented twice). After presenting each voice, the listeners were asked to decide whether the voice sounded high, moderate, or low in F0 (for the F0 condition), or fast, moderate, or slow in speech rate (for the speech rate condition), by pressing ‘1’, ‘2’, or ‘3’ on the numerical laptop keyboard. The listeners were asked to make this decision based on the voices that they hear in a real-world environment. The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life

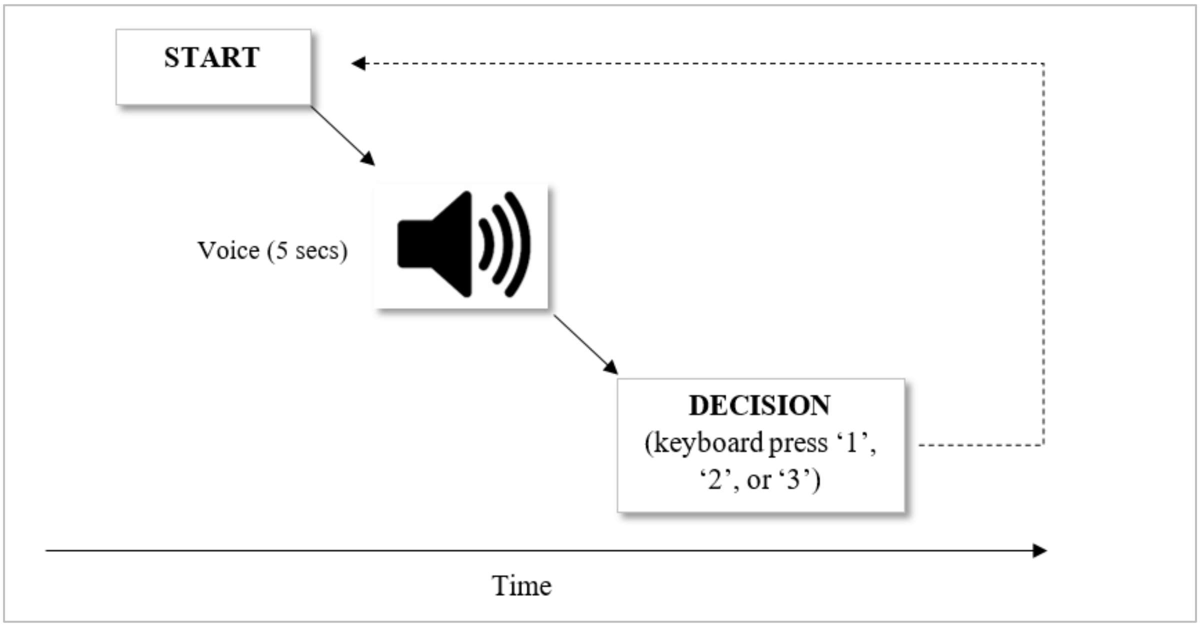


Figure 4.8: An illustration of the procedure in Experiment 1e.

4.3.5.3 Results and Discussion

4.3.5.3.1 Fundamental Frequency (F0)

Figure 4.9 presents the mean percentage of time that the target voices were categorised as being either high, moderate, or low in F0.

4.3.5.3.2 Speech Rate

Figure 4.10 presents the mean percentage of time that the target voices were categorised as being either fast, moderate, or slow in speech rate.

The results showed that for the F0 condition (Figure 8.1), the listeners assigned voices that were increased by 10% as being high in F0 at between 80% to 85% of the time, the original voices as being moderate in F0 between 75% to 80% of the time, and the voices that were decreased by 10% as being low in F0 between 85% to 90% of the time. For the speech rate condition (Figure 8.2), the listeners assigned voices that were increased by 20% as being fast in rate between 75% and 85% of the time, the original voices as being moderate in rate between 85% to 90% of the time, and the voices that were decreased by 20% as being slow in rate

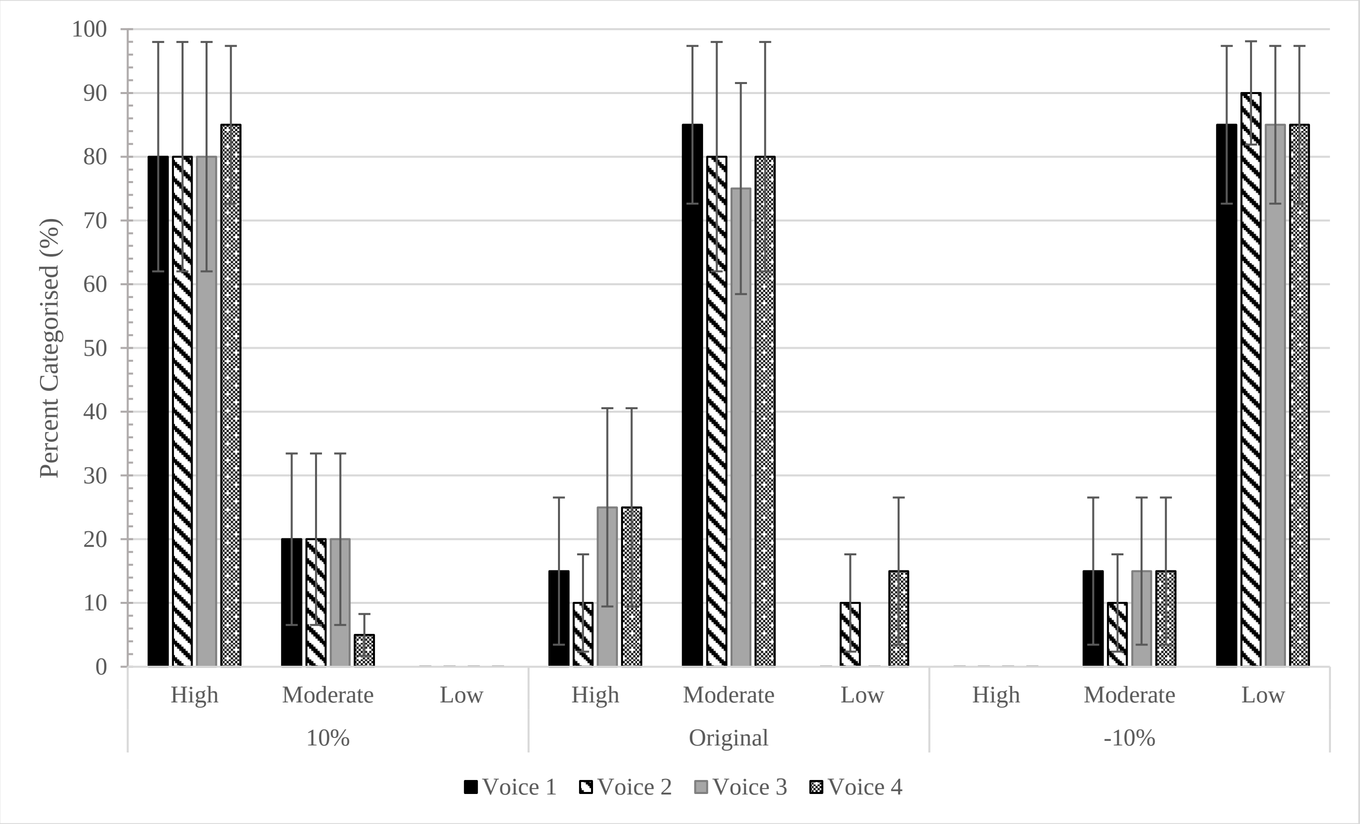


Figure 4.9: Mean percentage of time (%) the target voices were categorised as either high, moderate, or low in F0. Voice 1 and 2 depict male voices, and voice 3 and 4 depict female voices.

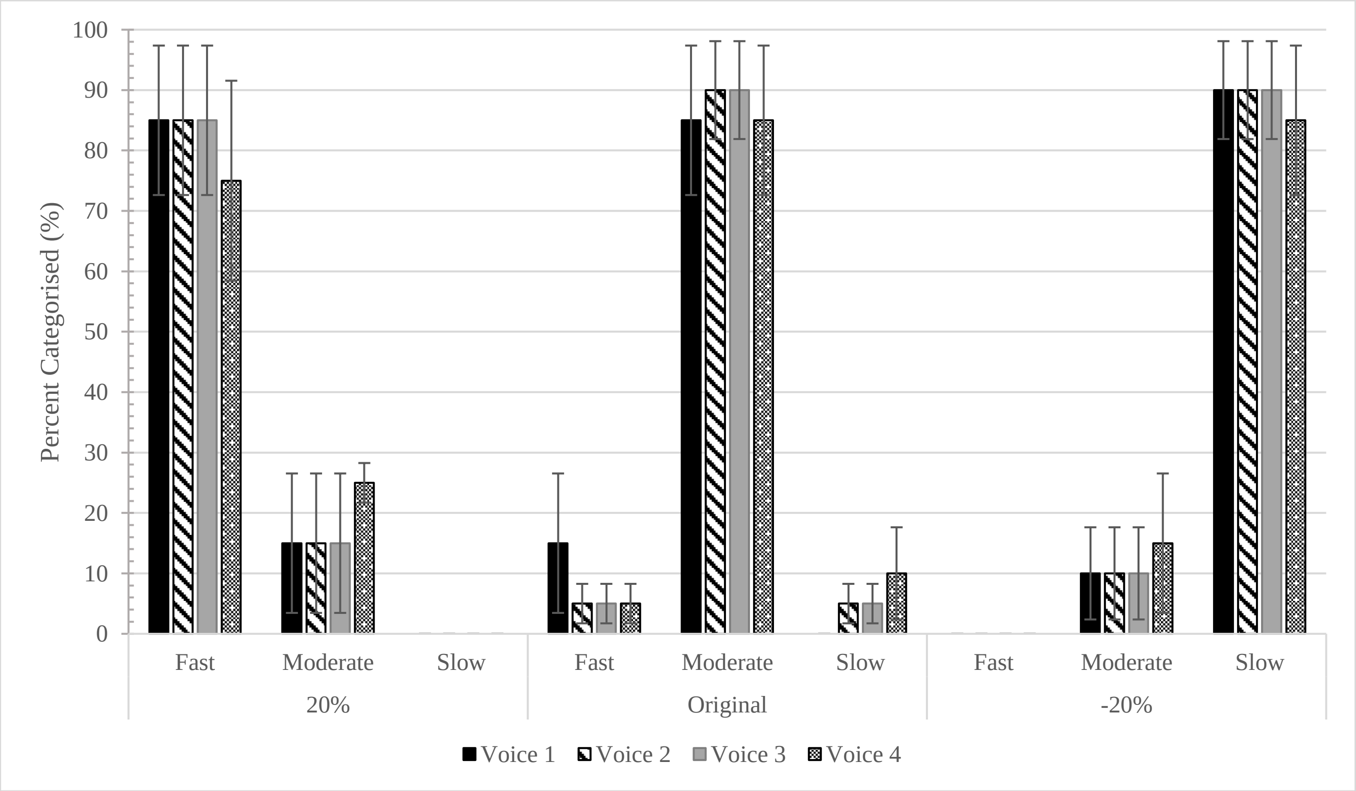


Figure 4.10: Mean percentage of time (%) the target voices were categorised as either fast, moderate, or slow in speech rate. Voice 1 and 2 depict male voices, and voice 3 and 4 depict female voices

between 85% to 90% of the time. Therefore, the target voices being used appeared to be approximately representative of high, moderate, and low F0, and fast, moderate, and slow speech rate voices heard in the real-world.

4.4 Summary Conclusions

- ❖ The text-to-speech synthesiser software Natural Reader 12.0 was used to generate synthesised voices for the experiments used in this thesis.
- ❖ Synthesised speech was used because of the need for precisely controlled stimuli that varied in fundamental frequency (F0) and speech rate, and to ensure that all of the voices were unfamiliar to the listeners.
- ❖ All of the voice samples were manipulated using Audacity software. This was chosen because it allowed the author to alter one characteristic (e.g., F0) whilst holding the other constant (e.g., speech rate).
- ❖ Measurements of fundamental frequency (F0) were calculated using Praat. Praat is a commonly used software application that can be used to precisely analyse speech sounds.
- ❖ Measurements of speech rate were calculated by dividing the total number of syllables in the utterance by the total number of seconds of the utterance, including pauses.
- ❖ Manipulations of all the voices in both fundamental frequency (F0) and speech rate were kept within the typical male and female F0 and speech rate range for voiced speech.
- ❖ Several extraneous variables were controlled for when choosing the stimuli for the experiments. These included ensuring the voices were unfamiliar to the listeners, all voices had a similar accent, the sentence spoken was non-emotive, and the duration of the sentences used were the same in length.

- Greater manipulations in fundamental frequency (F0) and speech rate increased the likelihood that the voices sounded less similar to the original voices.
- All of the voices used for the experiments were perceived as being different speakers (i.e., different identities).
- All of the voices used for the experiments sounded natural and were representative of real voices.
- The target voices used for Experiments 5a, 6, and 7 (high, moderate, and low F0) and Experiment 5b (fast, moderate, and slow speech rate) were representative of high, moderate, and low F0, and fast, moderate, and slow speech rates heard in the real-world.

CHAPTER 5. SPEAKER PERCEPTION: IDENTITY

5.1 Experiment 2: The Role of Fundamental Frequency (F0) and Speech Rate in Perceptions of Speaker Identity

5.1.1 Introduction

As previously described in Chapter 3 (Section 3.1.1.1), manipulations in F0, and to a lesser degree speech rate, have been investigated to determine the extent to which they affect perceptual judgements about the identity of the speaker (e.g., Kuwabara & Takagi, 1991; Lavner et al., 2000; Sell et al., 2015; Brown, 1981). The research tends to suggest that manipulations in F0 are more likely to change the identity of the speaker than manipulations in speech rate (e.g., Kuwabara & Takagi, 1991; Lavner et al., 2000; Sell et al., 2015; Brown, 1981). However, the evidence is somewhat limited and there are several methodological issues that make it difficult to determine whether manipulations in F0 or speech rate can change perceptual judgements about the identity of the speaker for unfamiliar voices, for both male and female speakers, and when complete sentences are used.

Therefore, the present study investigated whether manipulations in F0 or speech rate affect perceptual judgements of the identity of the speaker for a set of unfamiliar synthesised voices. A 2AFC perceptual discrimination paradigm was used in which listeners were presented with within voice pairs whereby one of the six original voices was paired with a manipulated version of that voice (i.e., increased or decreased in F0 or speech rate). The listeners task was to decide whether the pair of voices presented were the same identity or a different identity. The results of the experiment were analysed in two parts. In part one, the point of change (i.e., the point of subjective equality; PSE) at which listeners perceived the manipulated voices as being a different identity to the original voice was established. In part two, a one-way within-subjects ANOVA was used to determine the percentage of time listeners correctly identified the manipulated voice as being the same identity as the original voice (i.e., mean percentage accuracy).

5.1.1.1 Hypotheses

It was expected that manipulations in F0 would change the identity of the speaker. Specifically, greater manipulations in F0 would increase the likelihood that the voices were perceived as being different identities (i.e., different speakers). What some evidence exists, it is largely unknown what the effect of manipulations in speech rate have on the identity of the speaker. However, in line with the literature and the predictions for F0, it was also expected that manipulations in speech rate would change the identity of the speaker. Specifically, greater manipulations in speech rate would increase the likelihood that the voices were perceived as being different identities (i.e., different speakers).

5.1.2 Method

5.1.2.1 Participants

A total of 72 undergraduate students (36 males; 36 females) were recruited from Nottingham Trent University and received course credit for their participation. The ages of the participants ranged from 19 to 28 years old (M = 22.07 years, SD = 1.98 years). The inclusion criteria for the study required individuals to be between 18 and 30 years of age, have no known hearing deficits, have English as their first language, not undergone any musical training, and not heard the stimuli presented in the experiment before.

5.1.2.2 Stimuli and Materials

All six of the original voices and their subsequent manipulations (i.e., *for F0:* increased/decreased by 5% and 10%; *for Speech Rate:* increased/decreased by 5%, 10%, 15%, and 20%) were used for the experiments.

The speech samples were presented binaurally using Sony dynamic stereo headphones (Model No. MDR-V150). The experiment was run on a Sony Vaio laptop computer (Model No. SVF153B1YM) using PsychoPy version 1.7701 (Peirce, 2007) to control the presentation of the voices and collect participant responses.

5.1.2.3 Procedure

Using a perceptual discrimination paradigm, the participants were given a 2AFC (same/different key press) voice discrimination task. The order of presentation of the F0 and the speech rate conditions were counterbalanced across participants. As illustrated in Figure 5.1, in each trial, the stimuli were presented as within voice pairs, whereby an original voice was paired with itself or a modulated version (increased and decreased by 5% and 10% for the F0 condition, or increased and decreased by 5%, 10%, 15%, and 20% for the speech rate condition [2]), and presented in a random order. The text ‘Voice 1’ was visible in the middle of the screen while the first recording was playing, and the text ‘Voice 2’ was visible in the middle of the screen while the second voice was playing. Half of the time the original version was presented first, and half of the time the original version was presented second. There was a 1 second inter-stimulus interval between presentation of each voice. Following presentation of each trial, the participants had to indicate whether the two voices were the same person talking or a different person talking by pressing either the left or right keys on the laptop keyboard

[2] The manipulations were made based on the results from previous experimentation (refer to Chapter 4, Section 4.3.1) and allowed the author to determine whether the manipulated versions were discriminable from the original voice.

(counterbalanced across participants). For the F0 condition, there were 60 trials in total (five trials for each of the six voices, with each trial being presented twice). For the speech rate condition, there were 108 trials in total (nine trials for each of the six voices, with each trial being presented twice). The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life. Upon completion of the experiment, the participants were fully debriefed and thanked for their time and participation.

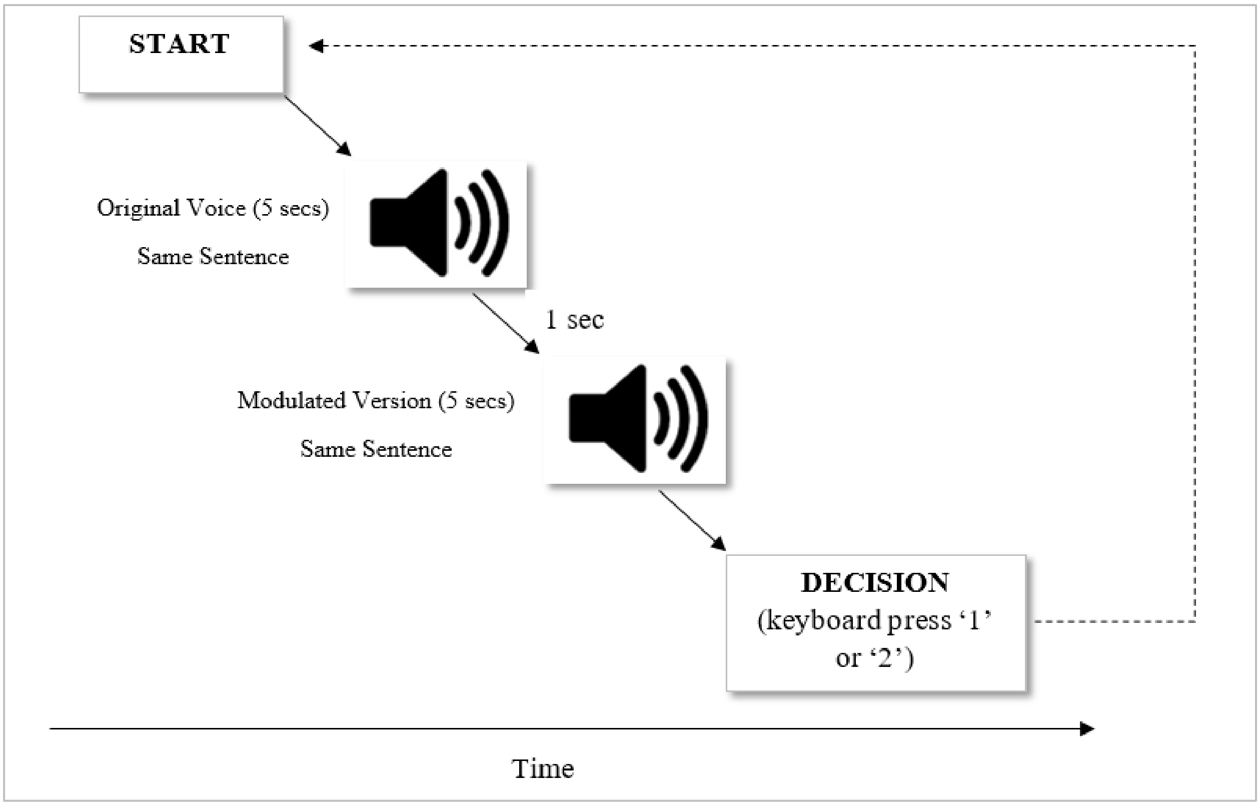


Figure 5.1: An illustration of the procedure in Experiment 2.

5.1.2.4 Analyses

The results were analysed in two parts. In part one, the results were analysed by plotting the percentage of time the listeners perceived the voices presented as being the same identity (i.e., the same speaker) as the original voice. This allowed the author to determine the difference threshold (i.e., the point of subjective equality; PSE) for the stimuli. A difference threshold is the point of intensity at which an observer can just detect a difference between two stimuli (taken at 50%, or chance level, for 2AFC tasks) (Gescheider, 1997). No difference is detected for stimuli with intensities below the threshold (i.e., the voices are perceived as being the same identity), whereas stimuli with intensities above the threshold are considered as being different (i.e., different identities). At the PSE, the observer perceives the two sounds to be the same. The data were collapsed across plus and minus manipulations for both the F0 and speech rate conditions owing to the comparative effect an increase and decrease in manipulation had on same speaker/different speaker ratings (refer to Appendix C1 and C2 for an illustration of this).

In part two, the results were analysed using one-way within-subjects ANOVA[3]. The within-subjects factor was distractor change (*for F0:* 0% (original voice), 5%, and 10%, *for speech rate:* 0% (original voice), 5%, 10%, 15%, ad 20%). The dependent variable measured was mean percentage accuracy (i.e., percentage of time listeners correctly identified the voice as being the 'same identity') [4]. Simple main effects were conducted using pairwise *t*-tests.

5.1.3 Results

5.1.3.1 Fundamental Frequency (F0)

5.1.3.1.1 Fundamental Frequency (F0): Part One

Figure 5.2 depicts the mean percentage of time listeners heard the voices presented as being the same identity (i.e., the same speaker) as the original voice, plotted separately for each of the six voices. The results suggest that listeners perceived there to be no difference in the identity of the speakers when an original voice was presented with itself (i.e., listeners correctly

[3] Note that data was collapsed across the six voices as there was no difference in the trend observed for each voice.
[4] For the original voice condition there were 72 participants. However, for the manipulated voice conditions, there were 144 participants in each. This is because the data was collapsed across plus and minus manipulations for the manipulated voice conditions.

identified these as being the same speaker with almost 100% accuracy). Listeners were also more likely to perceive the voices manipulated by 5% as being the same identity as the original voice (i.e., listeners correctly identified the voices as being the same speaker approximately 75% of the time). However, listeners were more likely to perceive the voices manipulated by 10% as being a different identity as the original voice (i.e., listeners correctly identified the voices as being the same speaker only 35% of the time). The average PSE for all six voices was also calculated at 8.6%, indicating that voices manipulated above this threshold are more likely to be perceived as a different speaker than the original voice, whereas voices manipulated below this threshold are more likely to be perceived as the same speaker as the original voice.

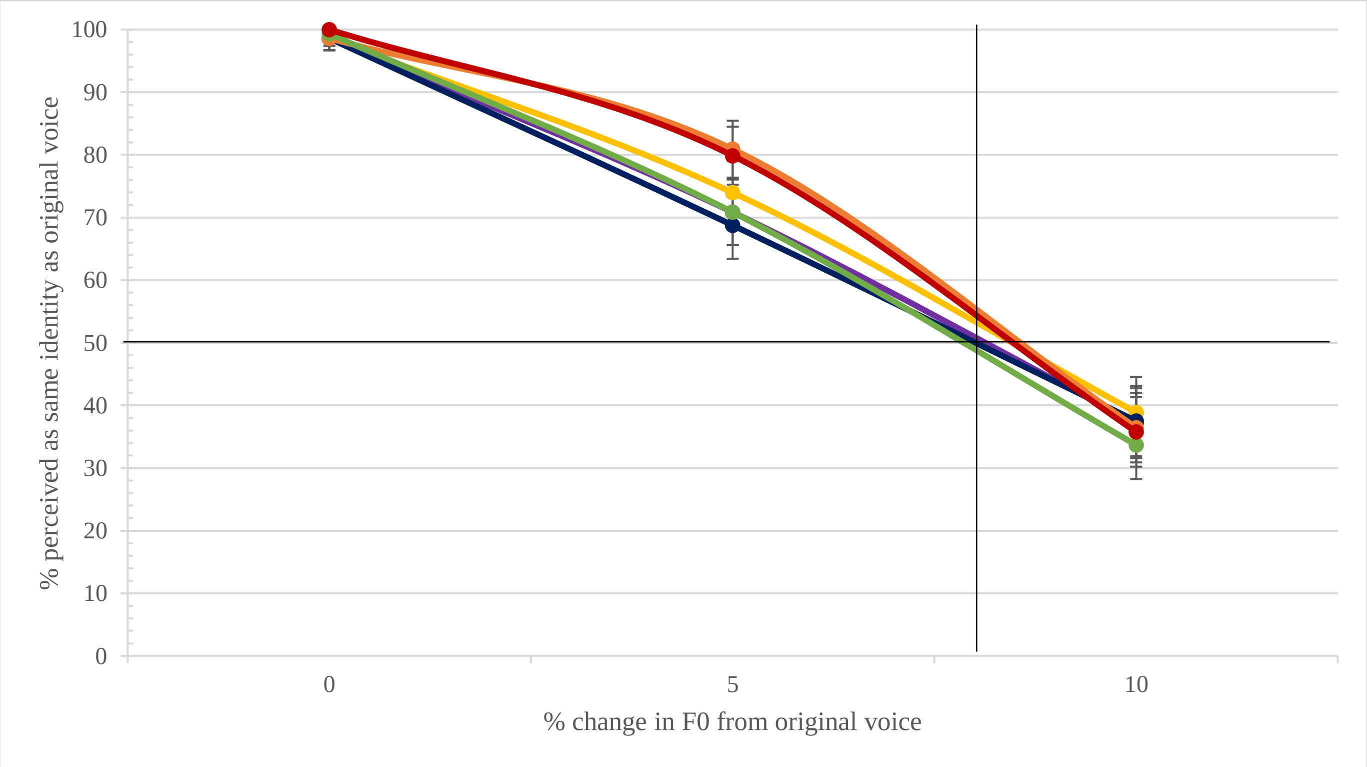


Figure 5.2: Line graph depicting the mean percentage of times listeners heard the manipulated versions of the voices as the same speaker (i.e., the same identity) as the original voice. Each of the six voices are depicted by a different line colour. Each of the three points for the different voices represents a version of that voice (from left to right; 0%, +/-5%, and +/-10%). Average PSE is 8.6%.

5.1.3.1.2 Fundamental Frequency (F0): Part Two

The overall mean accuracy scores were also entered into a within-subjects ANOVA. This revealed a significant main effect of distractor change, $F(2, 142) = 465.15$, $p = .001$, $\eta_g^2 = .79$. Listeners were significantly more accurate at judging the voices as being the same identity when the original voice was paired with itself ($M = 98.96$, $SD = 3.94$) compared to when the voices were manipulated by either 5% ($M = 74.19$, $SD = 17.10$) or 10% ($M = 36.57$, $SD = 19.75$). Listeners were also significantly more accurate at judging the voices as being the same identity when the original voices were manipulated by 5% ($M = 74.19$, $SD = 17.10$) compared to when they were manipulated by 10% ($M = 36.57$, $SD = 19.75$).

5.1.3.2 Speech Rate

5.1.3.2.1 Speech Rate: Part One

Figure 5.3 depicts the mean percentage of time listeners heard the voices presented as being the same identity (i.e., the same speaker) as the original voice, plotted separately for each of the six voices. The results suggest that listeners perceived there to be no difference in the identity of the speakers when an original voice was presented with itself (i.e., listeners correctly identified these as being the same speaker with almost 100% accuracy). Listeners were also more likely to perceive the manipulated versions of the voices (i.e., voices manipulated by 5%, 10%, 15%, and 20%) as being the same identity as the original voice. Therefore, whilst manipulations in speech rate did increase the uncertainty of the identity of the speaker by a small amount, listeners were still more likely to perceive the voices as being the same identity as the original voice 85% of the time or more.

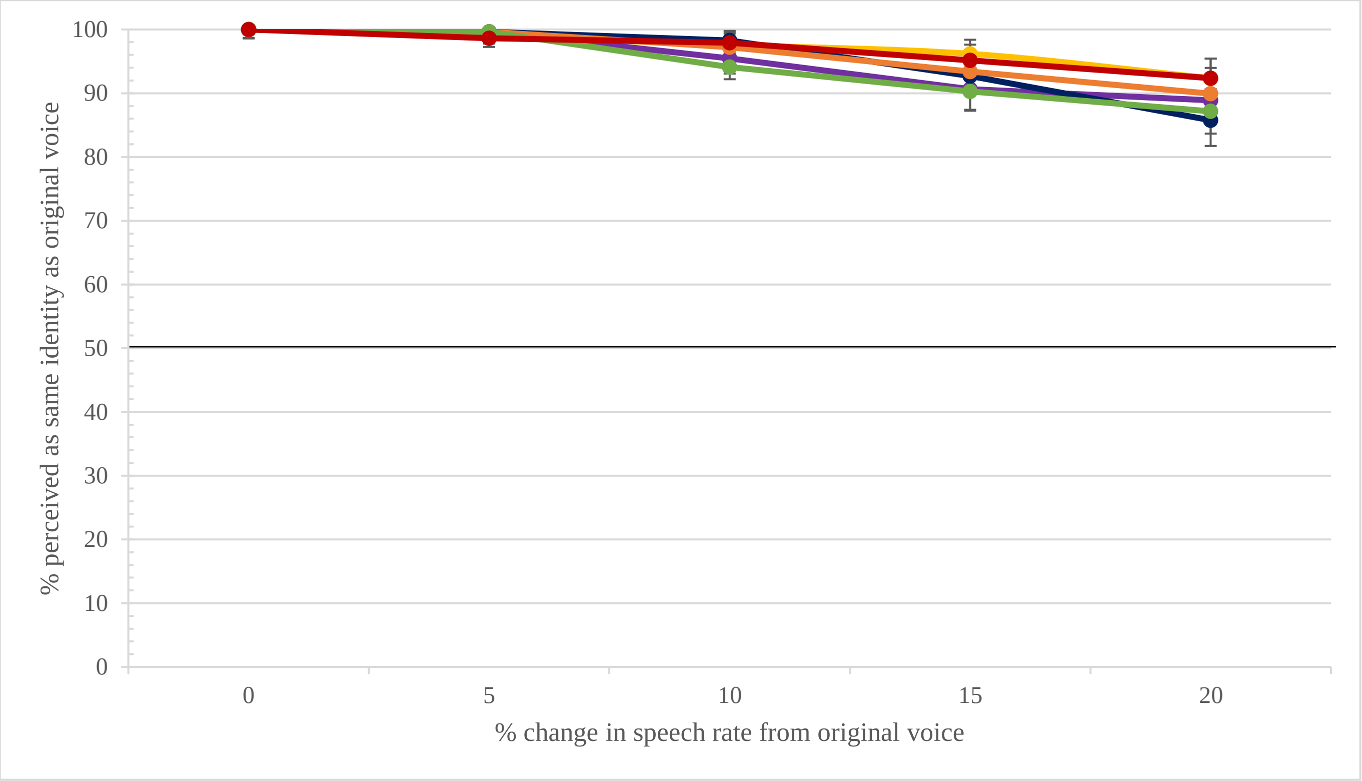


Figure 5.3: Line graph depicting the mean percentage of times listeners heard the manipulated versions of the voices as the same speaker (i.e., the same identity) as the original voice. Each of the six voices are depicted by a different line colour. Each of the five points for the different voices represents a version of that voice (from left to right; 0%, +/-5%, +/-10%, +/-15%, and +/-20%).

5.1.3.2.2 Speech Rate: Part Two

The mean accuracy scores were also entered in a one way within-subjects ANOVA. This revealed a significant main effect of distractor change, $F(2, 142) = 40.22$, $p = .001$, $\eta_g^2 = .73$. Listeners were significantly more accurate at judging the voices as being the same identity when the original voice was paired with itself ($M = 100.00$, $SD = 0.00$) compared to when the voices were manipulated by either 5% ($M = 99.36$, $SD = 1.66$), 10% ($M = 96.76$, $SD = 5.56$), 15% ($M = 93.06$, $SD = 9.49$), or 20% ($M = 89.41$, $SD = 11.55$).

5.1.4 Discussion

The present experiment investigated whether manipulations in F0 or speech rate affect perceptual judgements of the identity of the speaker for a set of unfamiliar synthesised voices. In line with the proposed hypothesis, greater manipulations in F0 increased the likelihood that voices would be perceived as a different identity (i.e., a different speaker) than the original voice. Greater manipulations in speech rate also increased the likelihood that voices would be perceived as being a different identity than the original voice. Listeners were more susceptible to making errors (i.e., perceiving the voices as being different identities) for manipulations in F0 than for manipulation in speech rate. The findings also showed that for both F0 and speech rate, listeners were more accurate at judging the voices as being the same identity when the original voice was paired with itself compared to when the original voice was paired with a manipulated version of the voice (i.e., a version manipulated by either 5% or 10% in F0, or 5%, 10%, 15%, or 20% in speech rate).

5.1.4.1 Fundamental Frequency (F0)

The results presented here offer additional support to the literature suggesting that manipulations in F0 are likely to affect the perceptual judgements of the identity of the speaker (e.g., Gaudrain et al., 2009; Kuwabara & Takagi, 1991; Lavner et al., 2000; Mathur et al., 2016; Sell et al., 2015). The finding that greater manipulations in F0 increased the likelihood that voice would be perceived as a different identity suggests that listeners do use F0 to help determine the identity of the speaker. This may be somewhat unsurprising given that F0 is strongly determined by the physiological and anatomical structures of the vocal tract (Fant, 1966), and therefore more directly related to speaker identity. The data reported here also show that manipulations in F0 affect perceptual judgements of the identity of the speaker similarly for both male and female voices. Previous work has often used only male voices (e.g., Gaudrain et al., 2009; Sell et al., 2015), making it difficult to determine whether the same acoustic cues are used to identify female speakers. The present experiment therefore suggests that listeners use F0 to determine identity for both male and female speakers.

The results of Experiment 2 showed that, contrary to previous findings, manipulations in F0 had a similar effect for all six synthesised voices. Previous research has shown that the ability to correctly identify a speaker is influenced differently by manipulations in F0 (Lavner, Gath, & Rosenhouse, 2000), suggesting that the susceptibility to misidentify a person is dependent on who is talking. The type of stimuli used in the present experiment may explain the difference in results. First, the current experiment used complete sentences to determine whether manipulations in F0 can affect perceptual judgements of the identity of the speaker. In contrast, research has typically used vowel sounds and nonsense words (e.g., Gaudrain et al., 2009; Lavner et al., 2000). Therefore, it is quite possible that the findings identified previously were related to the peculiarity of the stimuli used during experimentation. However, when complete sentences are used, as in the present experiment, any differences that were previously found to exist between speakers are no longer apparent. Second, the present experiment used unfamiliar voices, whereas previous research has often used familiar voices in their work (e.g., Kuwabara & Takagi, 1991; Lavner et al., 2000). It is possible that listeners are dependent on different cues of the voice when determining identity for familiar and unfamiliar voices. The findings of the present experiment suggest that listeners may be more reliant on F0 as a cue to the identity of the speaker for unfamiliar voices. Indeed, research has shown that the processes for identifying familiar and unfamiliar speakers are distinctly different (Yarmey et al., 2001). Further work would be required to confirm or disconfirm this explanation to the findings.

One important point to address is that the listeners in the present study were more likely to perceive the identity of the speaker as being different to the identity of the original voice at higher manipulations in F0. In contrast, studies that have used unfamiliar voices in their work have found listeners to consistently perform at a high level, and always above chance, when determining the identity of the speaker (Sell et al., 2015). Nevertheless, these studies made very small manipulations in F0. The present experiment has shown that small manipulations in F0 are unlikely to lead to significant changes in the perceived identity of the speaker. Therefore, it is probable that when larger manipulations in F0 are made, listeners will be likely to make more errors when determining the identity of the speaker. Again, further work would be required to confirm or disconfirm this explanation to the findings.

5.1.4.2 Speech Rate

Although manipulations in speech rate increased the uncertainty of the identity of the speaker by a small amount, listeners were highly robust to these changes and correctly perceived that the manipulated versions of the voice were the same identity as the original voice. This is perhaps unsurprising given that within-speaker variation is more typical in everyday situations for speech rate than it is for F0 (Mullenix et al., 2010; Stern et al., 2007). Furthermore, research has shown that changes in speech rate may be more likely to emphasise the intention of the spoken message rather than providing information about the identity of the speaker. Indeed, different rates of speech are often used in response to situational demands. For example, people have been found to speak slower when making public speeches (Gordon, Daneman, & Schneider, 2009). Changes in speech rate are also used to convey certain emotions. For example, increasing speech rate is likely to express excitement, anger, or fear (Siegman & Boyle, 1990). This is not to say that speech rate does not contain any identity information about the speaker. Indeed, slower or faster speaking styles are likely to be characteristic of certain speakers. However, such identity information may only be of value to the listener if the speaker is known to them (i.e., they are familiar with the speaker). For unfamiliar speakers however, such characteristic information will not be known to the listener, and so might not be used to determine the identity of the speaker. Further work would be required to conform of disconfirm the explanation to the findings.

5.1.4.3 Summary Conclusions

The results from the present experiment suggest that whilst greater manipulations in both F0 and speech rate increased uncertainty about the identity of the speaker, listeners are more robust to changes in speech rate than they are to changes in F0. Therefore, it can be concluded that F0 is more directly related to speaker identity than speech rate, and that listeners rely on F0 more than they do on speech rate when making decisions about the identity of the speaker. Experiment 3 (Chapter 6) will move on to consider the role of F0 and speech rate in perceptual judgements about the sex of the speaker.

CHAPTER 6. SPEAKER PERCEPTION: SEX

6.1 Experiment 3: The Role of Fundamental Frequency (F0) and Speech Rate in Perceptions of Speaker Sex

6.1.1 Introduction

Experiment 2 (Chapter 5) investigated whether manipulations in F0 or speech rate affect perceptual judgements of the identity of the speaker for a set of unfamiliar synthesised voices. The results showed that listeners were more susceptible to making errors (i.e., perceiving the voices as being different identities) for manipulations in F0 than for manipulations in speech rate. Greater manipulations in F0 increased the likelihood that voices would be perceived as a different identity than the original voice. The findings also showed that for both F0 and speech rate, listeners were more accurate at judging the voices as being the same identity when the original voices was paired with itself compared to when the original voice was paired with a manipulated version of the voice (i.e., a version manipulated by either 5% or 10% in F0, or 5%, 10%, 15%, or 20% in speech rate).

As previously discussed in Chapter 3 (Section 3.1.2.1 and 3.1.2.1), manipulations in F0, and to a lesser extent speech rate, have been investigated to determine the degree to which they affect perceptual judgements about the sex of the speaker (e.g., Assman et al., 2006; Coleman, 1971; Gelfer & Bennet, 2012; Gelfer & Mikos, 2005; Harnsberger et al., 2008; Hillenbrand & Clark, 2009; Whiteside, 1971). The research tends to suggest that manipulations in F0 are more likely to change the sex of the speaker than manipulations in speech rate. For F0, the literature tends to suggest a perceptual advantage for male speech where listeners have been found to hear talker sex somewhat more easily in male than in female voiced sounds

(Owren, Berkowitz, and Bachorowski, 2007). Specifically, the presence of critical features of maleness (i.e., low F0, low formants) almost certainly guarantees that the talker is an adult male. However, their absence does not unequivocally imply that the talker is an adult female (Owren, Berkowitz, and Bachorowski, 2007). Thus, manipulations in F0 are more likely to affect perceptions of speaker sex for female speakers than they are for male speakers. For speech rate, research has found that listeners believe females speak at a faster rate than males (Weirich & Simpson, 2014). However, studies tend to suggest that males actually have a faster speaking rate than females (Byrd, 1992; 1994). Nevertheless, in the main, research to determine the extent to which manipulations in speech rate affect perceptual judgements about the sex of the speaker is considerably lacking.

Therefore, the present study investigated whether manipulations in F0 or speech rate affect perceptual judgements of the sex of the speaker for a set of unfamiliar synthesised voices. The listeners were presented with one of the six original voices and the manipulated versions of the voices one at a time, and in a random order. Their task was to decide whether the voice they heard was a male voice or a female voice. The results of the experiment were analysed in two parts. In part one, the results were analysed by plotting the percentage of time the listeners perceived the voices presented as female. This allowed the author to determine the percentage of time listeners correctly perceived the voices as being either male (for the male voices) and female (for the female voices). In part two, a two-way within-subjects ANOVA was used to determine the percentage of time listeners correctly identified the voice as being male (for the male voices) and female (for the female voices).

6.1.1.1 Hypotheses

It was expected that manipulations in F0 would affect the perceptions of speaker sex. Specifically, decreasing F0 of female voices would lead to listeners being more likely to perceive the voices as male, whereas increasing F0 of male voices would lead to listeners being more likely to perceive the voices as female. Whilst some evidence exists, it is largely unknown what the effect of manipulations in speech rate have on perceptions of speaker sex. However, in line with the literature it was also expected that decreasing speech rate would lead to the listeners being more likely to perceive the voices as male, whereas increasing speech rate would lead to the listeners being more likely to perceive the voices as female.

6.1.2 Method

6.1.2.1 Participants

A total of 72 undergraduate students (36 males; 36 females) were recruited from Nottingham Trent University and they received course credit for their participation. The ages of the participants ranged from 19 to 30 years old (*M* = 25.04 years, *SD* = 2.16 years). The inclusion criteria for the study required individuals to be between 18 to 30 years of age, have no known hearing deficits, have English as their first language, not undergone any musical training, and had not heard the stimuli presented in the experiment before.

6.1.2.2 Stimuli and Materials

The materials and stimuli were identical to those used in Experiment 2 (Chapter 5). All six of the original voices and their subsequent manipulations (i.e., f*or F0:* increased/decreased by 5% and 10%; *for Speech Rate:* increased/decreased by 5%, 10%, 15%, and 20%) were used for the experiments.

6.1.2.3 Procedure

As illustrated in Figure 6.1, in each trial, the participants were presented with one of the six original voices or modulated versions of the voices (increased or decreased by 5% and 10% for the F0 condition, and increased or decreased by 5%, 10%, 15%, or 20% for the speech rate condition), and presented in a random order. The text ‘Voice’ was visible in the middle of the screen while the recording was playing. Following presentation of each voice, the participants were asked to decide whether the voice they heard was male or female by pressing either the left or right keys on the laptop keyboard (counterbalanced across participants). For the F0 condition, there were 60 trials in total (5 trials for each of the six voices, with each trial being presented twice). For the speech rate condition, there were 108 trials in total (nine trials for each of the six voices, with each trial being presented twice). The order of presentation of the F0 and the speech rate conditions were counterbalanced across participants. The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life. Upon completion of the experiment, the participants were fully debriefed and thanked for their time and participation.

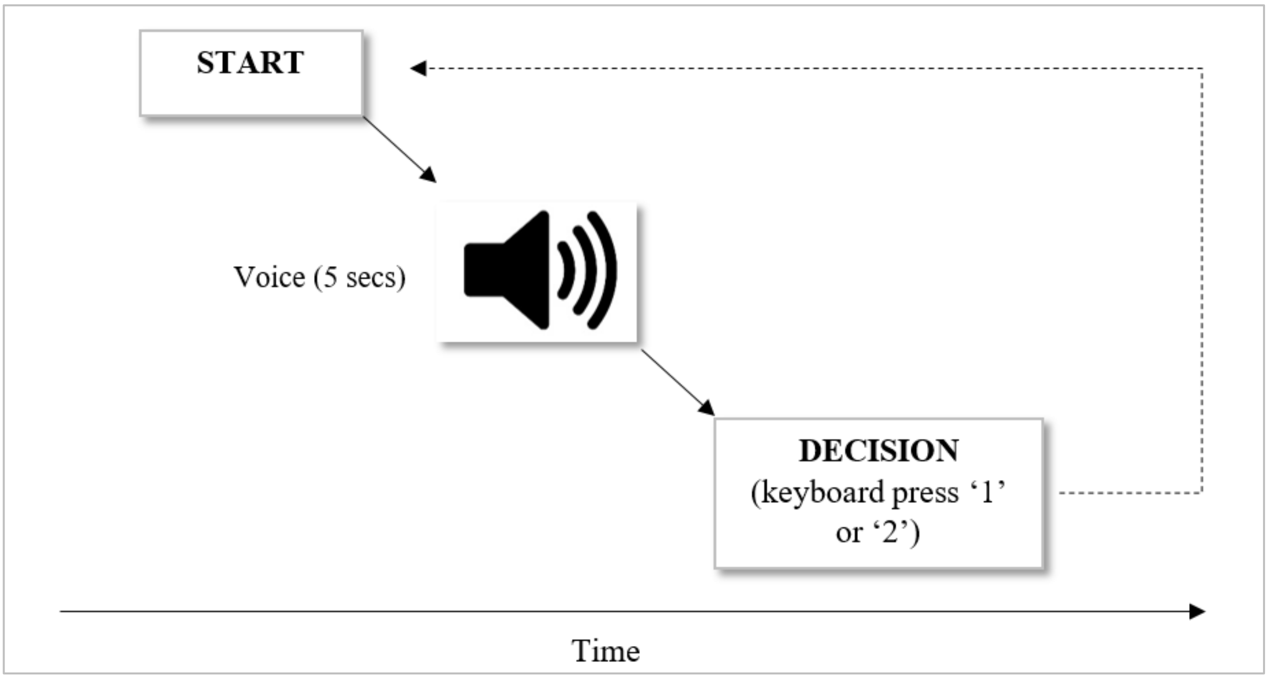


Figure 6.1: An illustration of the procedure used in Experiment 3.

6.1.2.4 Analyses

The results were analysed in two parts. In part one, the results were analysed by plotting the percentage of time the listeners perceived the voices presented as female. This allowed the author to determine the percentage of time listeners correctly perceived the voices as being either male (for the male voices) and female (for the female voices). Values greater than 50% signify that voices are more likely to be perceived as female, whereas values smaller than 50% signify that voices are more likely to be perceived as male. In part two, the results were analysed using a two-way within-subjects ANOVA[5]. The within-subjects factors were distractor change (*for F0:* -10%, -5%, 0% (original voice), 5%, and 10%, *for speech rate:* -20%, -15%, -10%, -5%, 0% (original voice), 5%, 10%, 15%, and 20%), and voice (voice 1, 2, and 3 *(male)*, and voices 4, 5, and 6 *(female)*). The dependant variable measured was mean percentage accuracy (i.e., percentage of time listeners correctly identified the voice as being male (for the male voices) and female (for the female voices). Simple main effects were conducted using pairwise *t*-tests.

6.1.3 Results

6.1.3.1 Fundamental Frequency (F0)

For F0, the findings will be presented as follows; part one will present the percentage of time the listeners perceived the voices as female, and part two will present the results of the two-way within-subjects ANOVA to determine the percentage of time listeners correctly identified the voice as being male (for male voices) and female (for female voices). Note, an additional experiment was also run to explore the findings from part one further. This additional experiment will subsequently be referred to as Part One (b) (see Section 6.1.3.1.2).

6.1.3.1.1 Fundamental Frequency (F0): Part One (a)

Figure 6.2 depicts the mean percentage of time that the listeners heard the speaker as female, plotted separately for each of the six voices. The data suggests that for the female voices (i.e., voice 4, voice, 5, and voice 6), decreasing F0 increased the likelihood that the

[5] Note that for Experiment 2 (Chapter 5), data was collapsed across the six voices as no difference in accuracy across the voices were observed. However, here, the ANOVA included voice as a separate variable due to the differences observed in accuracy for speaker sex (i.e., between male and female voices) in the F0 condition.

voices were perceived as male. This was particularly apparent for those female voices with a lower mean F0. F0 manipulations for female voices that fell within the typical F0 range for male voiced speech were either more likely to be perceived as male (i.e., voice 4), or perceived as male a substantial proportion of the time (i.e., voice 5 and voice 6). In contrast, for the male voices (i.e., voice 1, voice 2, and voice 3), manipulations in F0 did not increase the likelihood that the voices were perceived as female. Rather, listeners were accurate at perceiving the sex of all original and manipulated versions of the male voices.

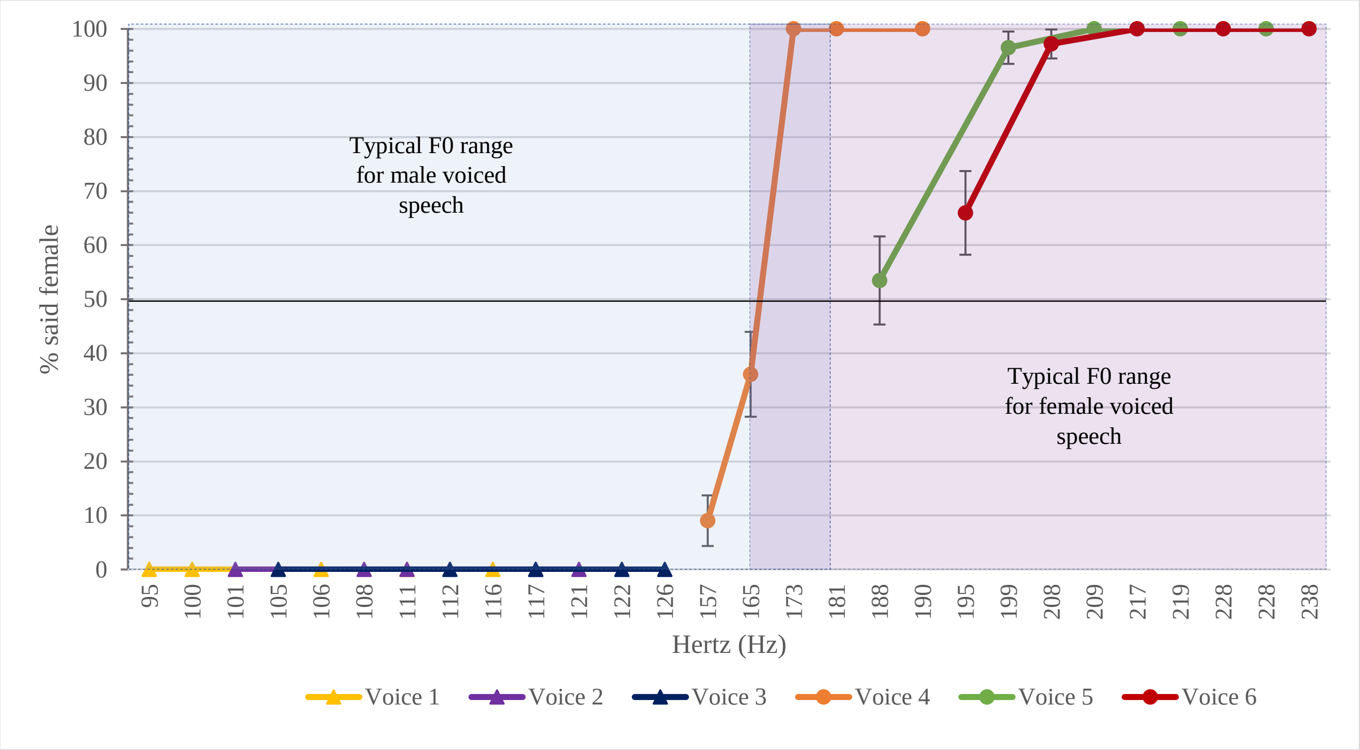


Figure 6.2: Line graph depicting the mean percentage of times listeners heard the original voices and the manipulated versions as female. Each of the six voices are depicted by a different line colour. The triangle symbol denotes a male voice, and a circle symbol denotes a female voice. Each of the five points for the different voices represents a version of that voice (from left to right; -10%, -5%, 0%, +5%, and +10%). The typical F0 range for both male and female voiced speech is shown. 95% confidence intervals are also shown.

6.1.3.1.2 Fundamental Frequency (F0): Part One (b)

To establish whether the findings for the female voices could be replicated, an additional experiment was re-run on 30 participants (15 males; 15 females) to determine whether a similar set of results would be obtained. The ages of the participants ranged from 19 to 27 years old (*M* = 24.63 years, *SD* = 1.57 years). However, this time in each trial, the participants were only presented with a voice that was decreased in F0 by 10%, and in a random order. Following presentation of each voice, the participants were asked to decide whether the voice they heard was male or female by pressing either the left or right keys on the laptop keyboard (counterbalanced across participants). There were 12 trials in total (with each voice being presented twice). An illustration of the procedure is shown in Figure 6.3.

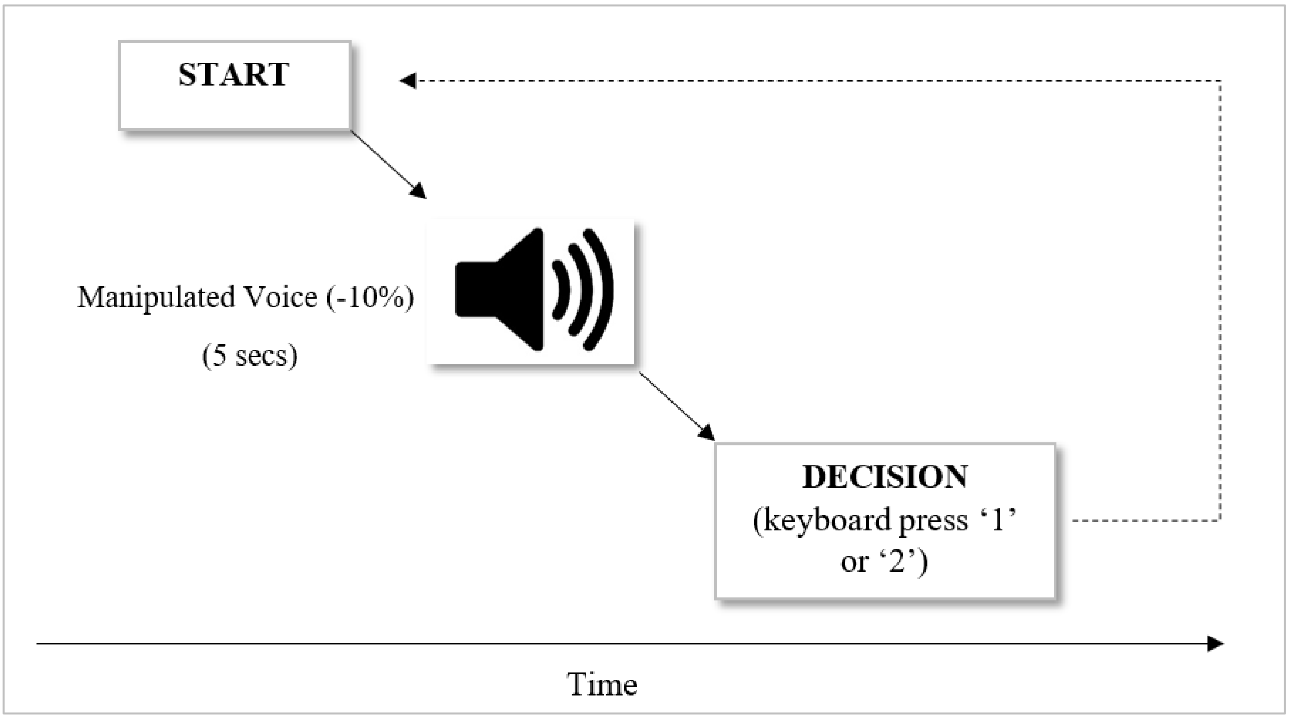


Figure 6.3: An illustration of the procedure used in Part One (b).

As illustrated in Figure 6.4, the results showed a very similar pattern to those observed previously. The data suggests that for female voices, decreasing F0 increased the likelihood that the voices were perceived as male. Again, this was particularly apparent for those female with a lower mean F0. F0 manipulations that fell within the typical F0 range for male voiced speech were either more likely to be perceived as male (i.e., voice 4), or perceived as male a substantial proportion of the time (i.e., voice 5 and voice 6). In contrast, for the male voices (i.e., voice 1, voice 2, and voice 3), manipulations in F0 did not increase the likelihood that the voices were perceived as female. Rather, listeners were accurate at perceiving the sex of all original and manipulated versions of the male voices. Therefore, it was concluded that the findings for the female voices could be replicated.

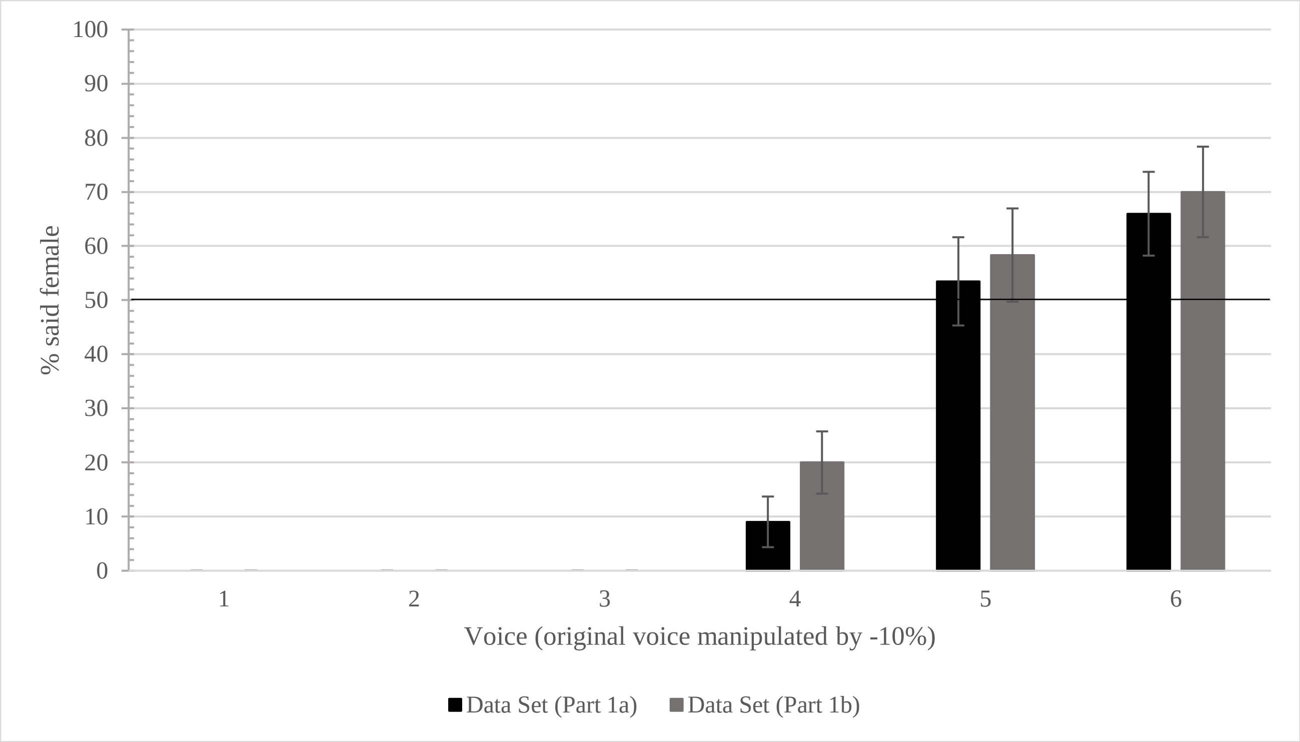


***Figure 6.4*:** Bar graph depicting the mean percentage of times listeners heard a voice that was decreased in F0 by 10% as female. Voice 1, 2, and 3 are male, and voice 4, 5, and 6 are female. The black bars represent the data obtained in part one (a), and the grey bars represent the data obtained in part one (b). 95% confidence intervals are also shown.

6.1.3.1.3 Fundamental Frequency (F0): Part Two

The overall mean accuracy scores were also entered into a two way within-subjects ANOVA. This revealed a significant main effect of voice, *F*(5, 355) = 323.51, *p* < .001, η_g^2 = .82. Listeners were more accurate at identifying speaker sex for male voices (*for voice 1: M* = 99.86, *SD* = 0.64*, for voice 2: M* = 100.00*, SD* = 0.00*, for voice 3: M* = 100.00*, SD* = 0.00) than they were for female voices (*for voice 4: M* = 69.03*, SD* = 13.37*, for voice 5: M* = 90.00*, SD* = 16.77*, for voice 6: M* = 92.64*, SD* = 16.61*), p* < .001. For female voices, listeners were also more accurate at identifying speaker sex for voice 5 (*M* = 90.00, *SD* = 16.77) and voice 6 (*M* = 92.64, *SD* = 16.61) than they were for voice 4 (*M* = 69.03, *SD* = 13.37), *p* < .001. However, no difference was observed between male voices (*for voice 1: M* = 99.86, *SD* = 0.64, *for voice 2: M* = 100.00, *SD* = 0.00, *for voice 3: M* = 100.00, *SD* = 0.00), *p* > 0.05. There was also a significant main effect of distractor change, *F*(4, 284) = 406.35, *p* < .001, η^2_g = .85. Listeners were significantly less accurate at identifying speaker sex for -10% manipulations (*M* = 71,41, *SD* = 40.71) than they were for -5% manipulations (*M* = 88.31, *SD* = 27.40), 0% manipulations (*M* = 100.00, *SD* = 0.00), 5% manipulations (*M* = 100.00, *SD* = 0.00), and 10% manipulations (*M* = 99.89, *SD* = 2.41), *p* < 0.01. Listeners were also significantly less accurate at identifying speaker sex for -5% manipulations (*M* = 88.31, *SD* = 27.40) than they were for 0% manipulations (*M* = 100.00, SD = 0.00), 5% manipulations (*M* = 100.00, *SD* = 0.00), and 10% manipulations (*M* = 99.89, SD = 2.41), *p* < 0.01. However, there was no difference in accuracy for 0% manipulations (*M* = 100.00, *SD* = 0.00) compared to 5% manipulations (*M* = 100.00, *SD* = 0.00) and 10% manipulations (*M* = 99.89, *SD* = 2.41), *p* > 0.05. There was also no difference in accuracy for 5% manipulations (*M* = 100.00, *SD* = 0.00) compared to 10% manipulations (*M* = 99.89, *SD* = 2.41), *p* > 0.05.

In addition to the main effects, there was also a significant interaction between voice and distractor change, *F*(20, 1420) = 162.73, *p* < .001, η^2_g = .70[6]. Listeners were significantly more accurate in determining speaker sex for the original female voice 4 (*M* = 100.00, *SD* = 0.00) compared to when the voice was decreased in F0 by 10% (*M* = 9.03, *SD* = 19.37) *t*(71) = -39.86, *p* < .001, *d* = 0.96, and by 5% (*M* = 36.11, *SD* = 28.11) *t*(71) = -19.28, *p* < .001, *d* = 0.85. Listeners were also significantly more accurate in determining speaker sex for the original female voice 5 (*M* = 100.00, *SD* = 0.00) compared to when the voice was decreased in F0 by 10% (*M* = 53.49, *SD* = 37.83), *t*(71) = -10.44, *p* < .001, *d* = 0.67. However, this was not the

[6] The author is interested in whether manipulations in F0 within-speakers can increase errors made identifying speaker sex. Therefore, only the simple main effects that are of direct interest are reported here (i.e., comparisons were made between the original voice and the manipulated versions of that voice). Simple main effects were not carried out for male voices because listeners were almost always accurate at correctly identifying speaker sex as male.

case when the voice was decreased in F0 by 5% (*M* = 96.53, *SD* = 17.46), *t*(71) = -1.69, *p* > .05, *d* = 0.14. A similar pattern of findings was observed for female voice 6, with listeners significantly more accurate in determining speaker sex for the original voice (*M* = 100.00, *SD* = 0.00) compared to when the voices was decreased in F0 by 10% (*M* = 65.97, *SD* = 38.30), *t*(71) = -7.54, *p* < .001, *d* = 0.53. However, this was not the case when the voice was decreased in F0 by 5 % (*M* = 97.22, *SD* = 11.53), *t*(71) = -2.04, *p* > .05, *d* = 0.17.

6.1.3.2 Speech Rate

6.1.3.2.1 Speech Rate: Part One

Figure 6.5 depicts the mean percentage of time that the listeners heard the speaker as female, plotted separately for each of the six voices. The data suggests that manipulations in speech rate did not affect perceptions of speaker sex. Rather, male voices were likely to be perceived as male, and female voices were likely to be perceived as female at all speeds with almost 100% accuracy.

6.1.3.2.2 Speech Rate: Part Two

The overall mean accuracy scores were also entered into a within-subjects ANOVA. This revealed a significant main effect of distractor, *F*(8, 568) = 6.73, *p* < .05, η_g^2 = .63, where listeners were significantly more accurate in determining speaker sex for the original voices (*M* = 99.88, *SD* = 0.98) compared to when the voice was decreased in rate by 20% (M = 98.38, *SD* = 3.60), *p* < .05. No other simple main effects were significant (*p* > 0.05). Neither the main effect of voice, *F*(5, 355) = 1.02, *p* > .05, η_g^2 = .01, nor the interaction effect between voice and distractor change, *F*(40, 2840) = 1.57, *p* > .05, η_g^2 = .02, was significant.

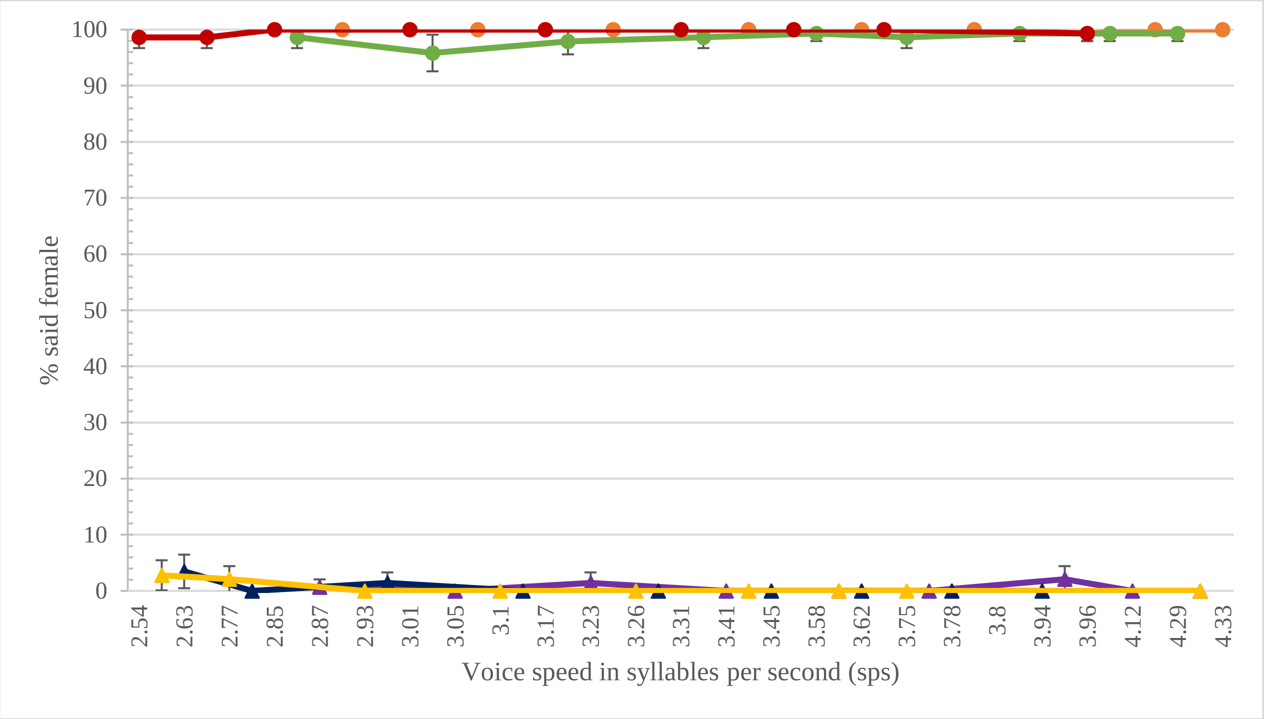


Figure 6.5: Line graph depicting the mean percentage of times listeners heard the original voices and the manipulated versions as female. Each of the six voices are depicted by a different line colour. The triangle symbol denotes a male voice, and a circle symbol denotes a female voice. Each of the nine points for the different voices represents a version of that voice (from left to right; -20%, -15%, -10%, -5%, 0%, +5%, +10%, +15%, and +20%). 95% confidence intervals are also shown.

6.1.4 Discussion

The present experiment investigated whether manipulations in F0 or speech rate affected perceptions of speaker sex. The results showed that when female voices were decreased in F0, the listeners were more likely to perceive the voices as male. This was particularly apparent for female voices with a lower overall mean F0, and for manipulations of voices that fell within the typical F0 male range for voiced speech. In contrast, for the male voices, manipulations in F0 did not increase the likelihood that male voices were perceived as female. Overall, listeners were more accurate perceiving speaker sex for male voices compared to female voices. Therefore, for female voices, the findings were in line with the original predictions made. However, for male voices, the finding did not support the original predictions made. For speech rate, the results showed that overall, listeners were accurate at perceiving speaker sex when voices were manipulated in speech rate. Nevertheless, listeners were less accurate in perceiving speaker sex when voices were decreased in rate by a large magnitude (i.e., 20%). Therefore, for speech rate, the findings did not support the original predictions made.

6.1.4.1 Fundamental Frequency (F0)

The results presented here offer additional support to the literature suggesting that manipulations in F0 are likely to affect perceptions of speaker sex (e.g., Assman et al., 2006; Coleman, 1971; Gelfer & Mikos, 2005; Hillenbrand & Clark, 2009; Whiteside, 1998). Female voices that were decreased in F0 and fell in the typical male F0 range for voiced speech were more likely to be perceived as male, provides evidence for a perceptual advantage for male speech where listeners hear talker sex more easily in male than in female voices (Owren, Berkowitz, and Bachorowski, 2007). This is further supported by the finding that male voices increased in F0 did not change perceptions of speaker sex (i.e., the listeners did perceive these voices as female). However, it should be noted that when male voices were increased in F0, they did not fall in either the gender ambiguous range or the typical female F0 range for voiced speech. Therefore, it is difficult to determine whether male voices that fell within the typical female F0 range for voiced speech would increase the likelihood that these would be perceived as female. Nevertheless, given that the present findings showed that increasing F0 did not change the perceived sex of the speaker for male voices at all, it is unlikely that listeners would have perceived the voices as female, or at least at a greater percentage of the time than they would have perceived the voices as male.

Further support for the male advantage hypothesis (Owren, Berkowitz, and Bachorowski, 2007 comes from the finding that even small manipulations (i.e., 5%) in F0 for female voices increased the likelihood that voices were perceived as male, even though they remained in the typical female F0 range for voiced speech. Nevertheless, observations of vowel formant frequencies for these voices fell in the typical male range (refer to Chapter 4, Section 4.1.1.1 for further information). Such findings therefore offer support to the existing literature that suggests both F0 and formant frequencies are used to determine the sex of the speaker, especially in situations of uncertainty (Assman et al., 2006; Coleman, 1971; Gelfer & Mikos, 2005 Hillenbrand & Clark, 2009). What's more, this is in line with previous findings that have shown that when female F0 is paired with male formant frequencies, listeners are more likely to identify the voice as male rather than female (Coleman, 1971). However, it is important to note that the present findings suggest that whilst listeners did perceive voices with lower formant frequencies as male, they were still more likely to perceive the voices as female, lending support to the suggestion that F0 is the more robust cue in determining speaker sex (Coleman, 1971).

6.1.4.2 Speech Rate

For speech rate, listeners accurately identified the sex of the speaker when voices were increased and decreased in speech rate. This is perhaps unsurprising given that male and female speakers do not differ considerably in their rate of speech, and so listeners do not use this as a cue to determine speaker sex. However, given that the stereotypical opinion is for females to have a faster speaking rate than males (Weirich & Simpson, 2014), it is possible that manipulations in speech rate may have affected perceptions of speaker sex. One possible explanation for such findings is that manipulations in speech rate did not change the F0 of the speaker. Research has typically shown that faster speaking rates are also perceived as having a higher F0 (Bond & Feldstein, 1982). Therefore, in the present experiment, manipulations in speech rate may not have affected perceptions of speaker sex because the voices used were not typical of those that are likely to change the perceived sex of the speaker.

Listeners were less accurate in perceiving speaker sex when voices were decreased in speech rate by a large magnitude (i.e., 20%). For female voices, these findings are consistent with both the original predictions made and with the existing literature suggesting that the stereotypical opinion is for females to speak at a faster rate than males (e.g., Weirich & Simpson, 2014). For male voices, whilst these findings are inconsistent with the original predictions made, studies tend to suggest that males actually have a faster speech rate than females (Byrd, 1992, 1994), and thus offering some support to the actual differences observed between speech rate for male and female voices.

Another explanation for the finding that listeners were less accurate in perceiving speaker sex when voices were decreased in speech rate by a large magnitude (i.e., 20%), could be explained by how familiar listeners were with the slower rate voices used in the experiment. In natural speech, a person speaking more slowly is likely to be more hesitant, making more silent pauses or filled pauses (e.g., *um, er*). In the present experiment, decreasing speech rate did affect the rate of continuous production, nevertheless, it did not result in any increased pauses or hesitations of any kind. Whilst the manipulations made are consistent with those in previous research (e.g., Assman et al., 2006; Coleman, 1971; Gelfer & Bennet, 2012; Gelfer & Mikos, 2005; Harnsberger et al., 2008; Hillenbrand & Clark, 2009; Whiteside, 1971), the speech samples used in the present experiment may not be an entirely natural rendition of slower speech, or at least of a type that listeners most typically hear. Thus, listeners may have been likely to make more errors for slow rate speech because the voices were not representative of those often heard in the real-world.

6.1.4.3 Summary Conclusions

The results from the present experiment suggest that manipulations in F0 are more likely to increase uncertainty about speaker sex than manipulations in speech rate. Voices that are decreased in speech rate do increase the uncertainty of speaker sex. However, overall, listeners are accurate at determining speaker sex when voices are manipulated in speech rate. For F0, listeners are accurate in perceiving speaker sex when F0 is increased for both male and female voices. Nevertheless, decreasing F0 of female voices increased the uncertainty of speaker sex (i.e., voices were more likely to be perceived as male rather than female). Consequently, it is likely that male cues are more salient than female cues for determining speaker sex. It can be concluded that F0 is more directly related to speaker sex than speech rate, and that listeners rely on F0 more than they do on speech rate when making decisions about the sex of the speaker. Experiment 4 (Chapter 7) will move on to consider the role of F0 and speech rate in perceptual judgements about the age of the speaker.

CHAPTER 7. SPEAKER PERCEPTION: AGE

7.1 Experiment 4: The Role of Fundamental Frequency (F0) and Speech Rate in Perceptions of Speaker Age

7.1.1 Introduction

The experiments carried out so far have demonstrated that manipulations in F0 are likely to change both the perceived identity and the sex of the speaker. Experiment 2 (Chapter 5) showed that listeners were more susceptible to making errors (i.e., perceiving the voices as being different identities) for manipulations in F0 than for manipulations in speech rate. Greater manipulations in F0 increased the likelihood that voices would be perceived as a different identity than the original voice. The findings also showed that for both F0 and speech rate, listeners were more accurate at judging the voices as being the same identity when the original voices was paired with itself compared to when the original voice was paired with a manipulated version of the voice (i.e., a version manipulated by either 5% or 10% in F0, or 5%, 10%, 15%, or 20% in speech rate). Experiment 3 (Chapter 6) showed that manipulations in F0 are more likely to increase uncertainty about speaker sex than manipulations in speech rate. Overall, listeners were accurate at determining speaker sex when voices were manipulated in speech rate. However, voices decreased in speech rate increased the uncertainty of speaker sex. For F0, listeners were accurate in perceiving speaker sex when F0 was increased for both male and female voices. Nevertheless, decreasing F0 of female voices increased the uncertainty of speaker sex (i.e., voices were more likely to be perceived as male rather than female).

As previously discussed in Chapter 3 (Section 3.1.3.3), manipulations in both F0 and speech rate have been investigated to determine the extent to which they affect perception of the age of the speaker (e.g., Linville & Fisher, 1985; Shipp et al., 1992; Shrivastav et al., 2003; Smith & Patterson, 2005; Waller & Eriksson, 2016). The research tends to suggest that F0 and speech rate are both important in estimating perceptions of speaker age. For F0, research has shown that manipulations in F0 are unlikely to influence perceptions of age (Braun & Rietveld, 1995), whereas others have found estimates of perceived age to be lower (i.e., younger) when F0 is decreased (Hollien et al., 2003). Whilst this is in line with the typical pattern observed over the lifespan for male voices (i.e., F0 of male speakers continues to fall before it begins to rise again at around 60 to 70 years of age), this is particularly surprising for female voices given that F0 typically decreases in female speakers as age increases. Thus, for F0, estimations of speaker age may be based on the stereotyping of vocal cues rather than the actual change that occurs in the F0 of the speaker. For speech rate, estimates of perceived age are often lower (i.e., younger) when speech rate is increased and higher (i.e., older) when speech rate is decreased (Hollien et al., 2003). Taken together, the results tend to suggest that whilst listeners do use F0 to determine the age of the speaker, estimates of speaker age may be influenced more strongly by manipulations in speech rate than manipulations in F0, although the findings are still somewhat inconclusive.

Therefore, the present study investigated whether manipulations in F0 or speech rate affect the perceptual judgements of the age of the speaker for a set of unfamiliar synthesised voices. The listeners were presented with one of the six original voices and the manipulated versions of the voices one at a time, and in a random order. Their task was to freely estimate the age of the speaker. The results of the experiment were analysed in two parts. In part one, the results were analysed by plotting listeners judgements about the voices mean age (in years). This allowed the author to determine whether manipulations in F0 or speech rate affect perceptions of speaker age. In part two, a one-way ANOVA was used to determine the whether the mean age of the manipulated voices were different from the mean age of the original voices.

7.1.1.1 Hypotheses

It was expected that manipulations in F0 and speech rate would affect perceptions of speaker sex. Increasing F0 would lead to listeners perceiving the voices as younger than the original voices, whereas decreasing F0 would lead to listeners perceiving the voices as older than the original voices. Increasing speech rate would lead to listeners perceiving the voices as younger than the original voices, whereas decreasing speech rate would lead to listeners perceiving the voices as older than the original voices.

7.1.2 Method

7.1.2.1 Participants

A total of 72 undergraduate students (36 males; 36 females) were recruited from Nottingham Trent University and they received course credit for their participation. The ages of the participants ranged from 18 to 26 years old (*M* = 21.78 years, *SD* = 2.47 years). The inclusion criteria for the study required individuals to be between 18 to 30 years of age, have no known hearing deficits, have English as their first language, and not undergone any musical training.

7.1.2.2 Stimuli and Materials

The materials and stimuli were identical to those used in Experiment 2 (Chapter 5) and Experiment 3 (Chapter 6). All six of the original voices and their subsequent manipulations (i.e., *for F0:* increased/decreased by 5% and 10%; *for Speech Rate:* increased/decreased by 5%, 10%, 15%, and 20%) were used for the experiments.

7.1.2.3 Procedure

The procedure was identical to that in Experiment 3 (Chapter 6). As illustrated in Figure 7.1, in each trial, the participants were presented with one of the six original voices or modulated versions of the voices (increased or decreased by 5% and 10% for the F0 condition, and increased or decreased by 5%, 10%, 15%, or 20% for the speech rate condition), and presented in a random order. The text ‘Voice’ was visible in the middle of the screen while the recording was playing. Following presentation of each voice, the participants were asked to estimate the age of the voice by keying in the age using the number keys on the laptop keyboard. For the F0 condition, there were 60 trials in total (5 trials for each of the six voices, with each trial being presented twice). For the speech rate condition, there were 108 trials in total (nine trials for each of the six voices, with each trial being presented twice). The order of presentation of the F0 and the speech rate conditions were counterbalanced across participants. The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life. Upon completion of the experiment, the participants were fully debriefed and thanked for their time.

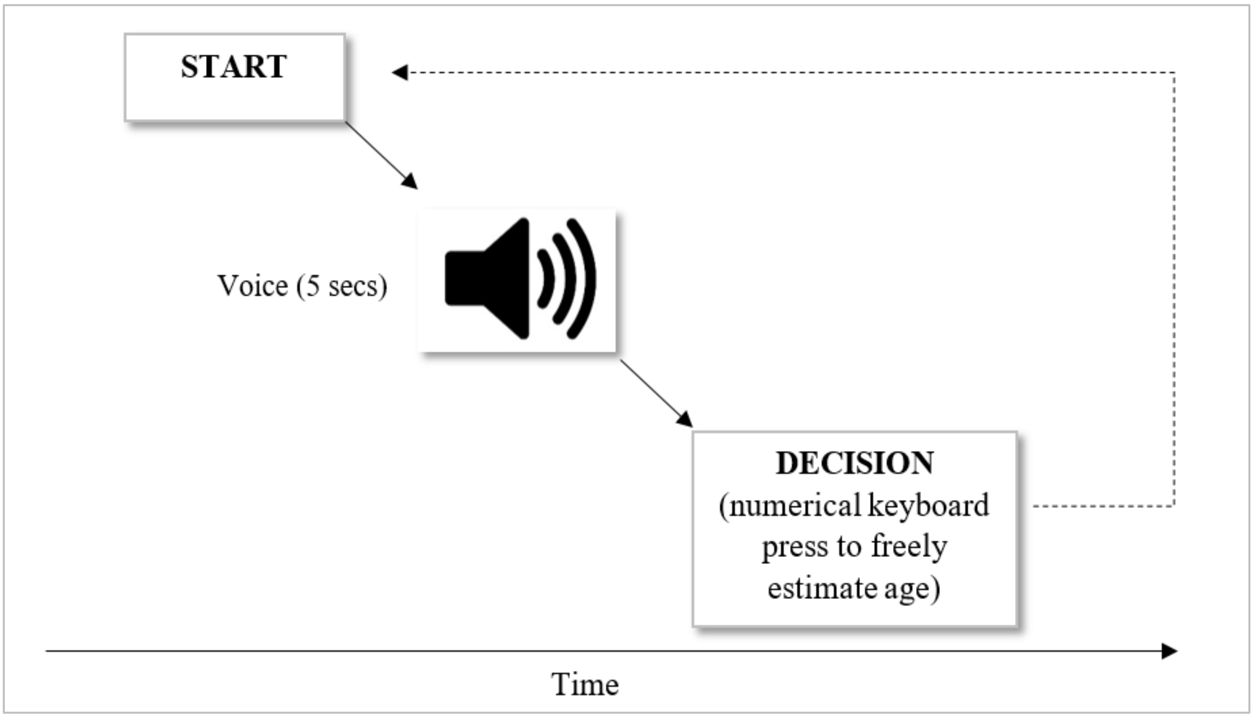


Figure 7.1: An illustration of the procedure in Experiment 4.

7.1.2.4 Analyses

The results were analysed in two parts. In part one, the results were analysed by plotting listeners judgements about the mean age (in years) for the six original voices and the manipulated versions of the voices. This allowed the author to determine whether manipulations in F0 or speech rate affect perceptions of speaker age. In part two, the results were analysed using a one-way within-subjects ANOVA[7]. The within-subjects factor was distractor change (*for F0:* -10%, -5%, 0% (original voice), 5%, and 10%, *for speech rate:* -20%, -15%, -10%, -5%, 0% (original voice), 5%, 10%, 15%, and 20%). The dependant variable measured was mean age (in years). Simple main effects were conducted using pairwise *t*-tests.

7.1.3 Results

7.1.3.1 Fundamental Frequency (F0)

7.1.3.1.1 Fundamental Frequency (F0): Part One

Figure 7.2 depicts the mean age (in years) for the original voices and the manipulated versions of the voices, plotted separately for each of the six voices. The data suggests that for both male and female speakers, manipulations in F0 lead to listeners perceiving the voices as being different ages than the original voice. Greater manipulations in F0 (i.e., +10% and -10%) lead to greater changes in the perceived age of the speakers, whereas smaller manipulations in F0 (i.e., +5% and -5%) lead to smaller changes in the perceived age of the speakers. Increasing F0 lead to listeners perceiving the voices as being younger in age than the original voices, whereas decreasing F0 lead to listeners perceiving the voices as being older in age than the original voices.

[7] Note that the data was collapsed across the six voices as there was no difference in the trend observed for each voice.

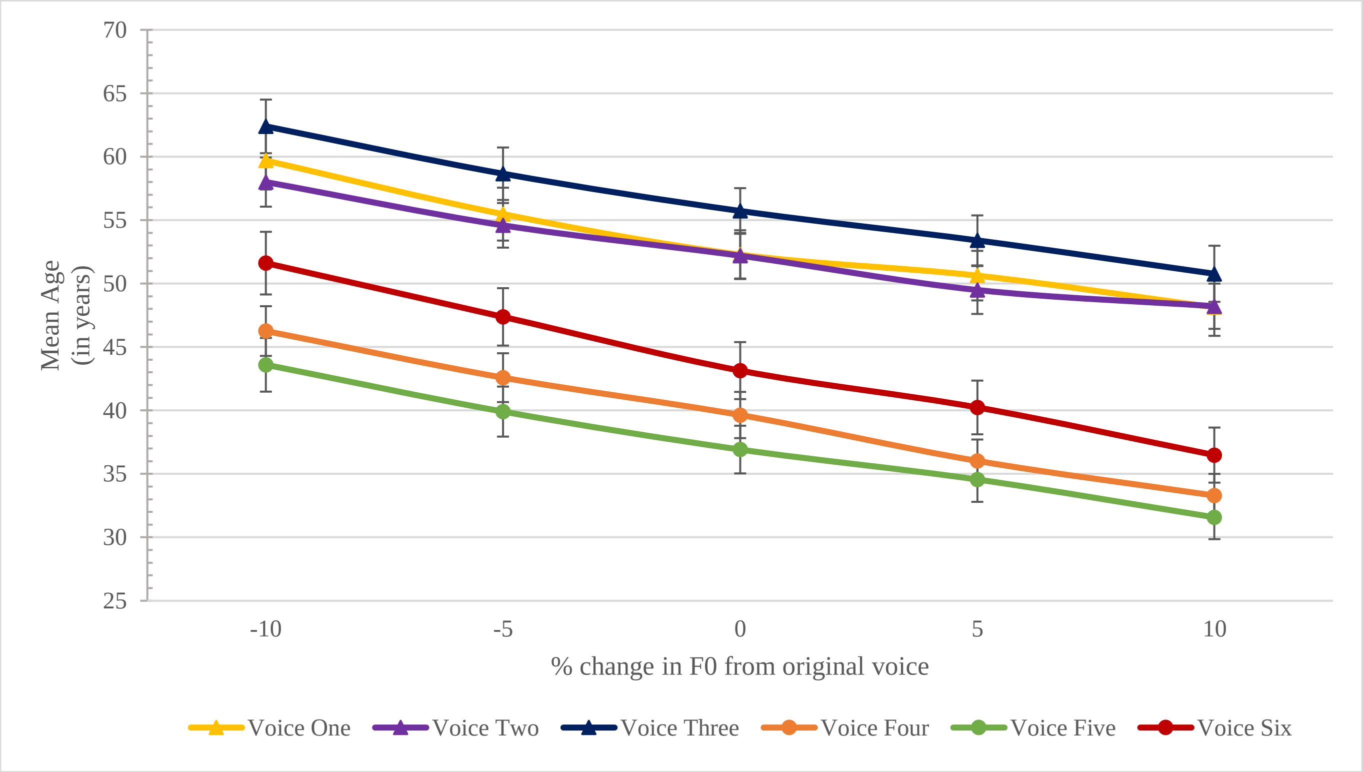


Figure 7.2: Line graph depicting perceived mean age (in years) for the original voices and the manipulated versions of the voices. Each of the six voices are depicted by a different line colour. The triangle symbol denotes a male voice, and a circle symbol denotes a female voice. Each of the five points for the different voices represents a version of that voice (from left to right; -10%, -5%, 0%, +5%, and +10%). 95% confidence intervals are also shown.

7.1.3.1.2 Fundamental Frequency (F0): Part Two

The overall mean accuracy scores were entered into a one way within-subjects ANOVA. This revealed a significant distractor change, $F(1, 143) = 575.12$, $p < .001$, , $\eta_g^2 = .80$[8]. Voices were estimated as being significantly older when they were manipulated by -10% (M = 53.59, SD = 6.54) compared to the original voices (M = 46.65, SD = 5.53), $t(143) = -22.03$, $p < .001$, $d = 1.15$. Voices were also estimated as being significantly older when they were manipulated by -5% (M = 49.77, SD = 5.85) compared to the original voices (M = 46.65,

[8] The author is interested in whether manipulations in F0 within-speakers can increase errors made in perceptions of speaker age. Therefore, only the simple main effects that are of direct interest are reported here (i.e., comparisons were made between the original voice and the manipulated version of that voice).

SD = 5.53), *t*(143) = -14.43, *p* < .001, *d* = 0.55. In contrast, voices were estimated as being significantly younger when voices they were manipulated by 5% (*M* = 44.06, *SD* = 5.63) compared to the original voices (*M* = 46.65, *SD* = 5.53), *t*(143) = 21.63, *p* < .001, *d* = 0.46. Voices were also estimated as being significantly younger when they were manipulated by 10% (*M* = 49.77, *SD* = 5.85) compared to the original voices (*M* = 46.65, *SD* = 5.53), *t*(143) = 14.22, *p* < .001, *d* = 0.55.

A paired samples *t*-test was also conducted to compare whether there was any difference in age estimates between male and female voices. There was a significant difference in age estimates for male and female voices. Specifically, female voices were estimated as being significantly younger (*M* = 40.21, *SD* = 6.13) compared to male voices (*M* = 54.00, *SD* = 6.32), *t*(143) = 28.67, *p* < .001, *d* = 0.74.

7.1.3.2 Speech Rate

7.1.3.2.1 Speech Rate: Part One

Figure 7.3 depicts the mean age (in years) for the original and the manipulated versions, plotted separately for each of the six voices. The data suggests that for both male and female speakers, manipulations in speech rate lead to listeners perceiving the voices as being different ages than the original voice. Greater manipulations in speech rate (i.e., +20% and -20%) lead to greater changes in the perceived age of the speakers, whereas smaller manipulations in speech rate (i.e., +5% and -5%) lead to smaller changes in the perceived age of the speakers. Increasing speech rate lead to listeners perceiving the voices as being younger in age than the original voices, whereas decreasing speech rate lead to listeners perceiving the voices as being older in age than the original voices.

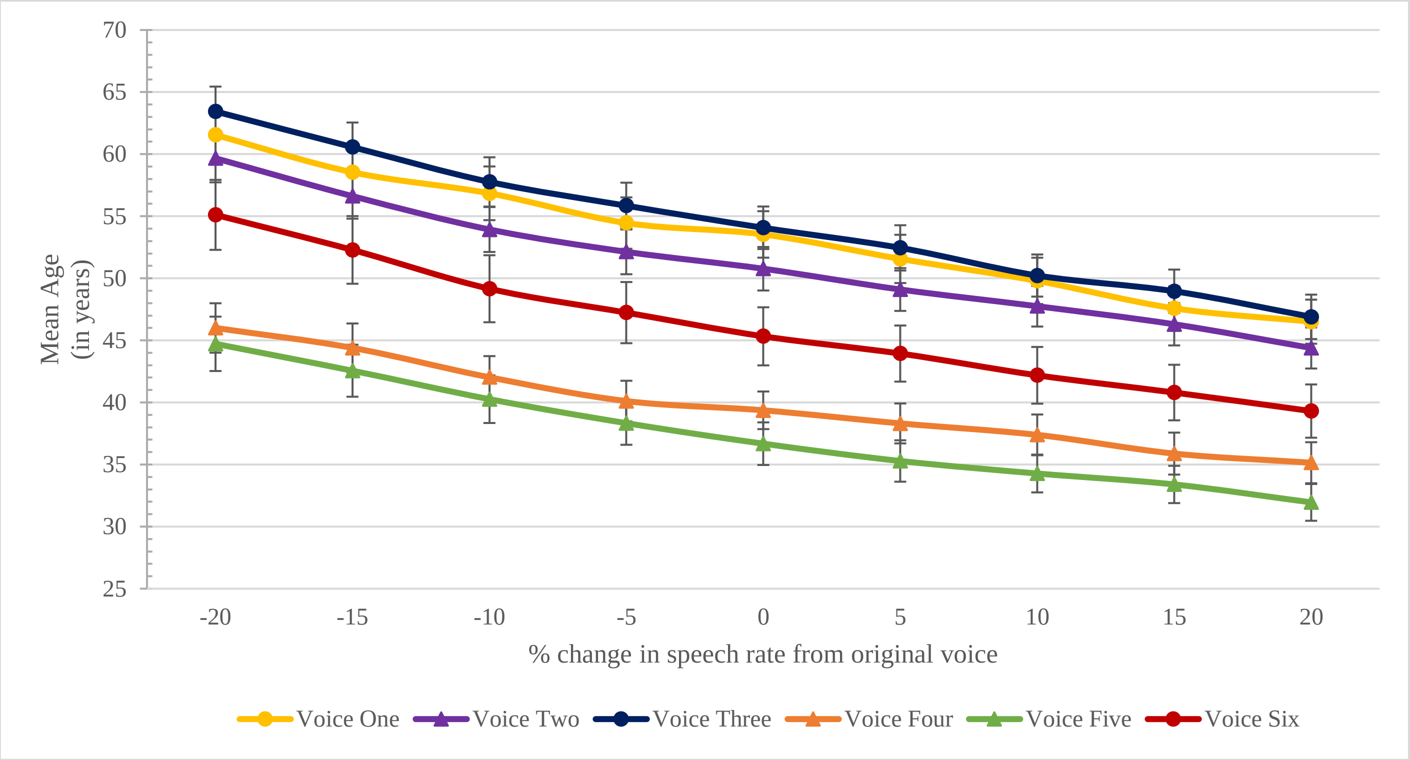


Figure 7.3: Line graph depicting perceived mean age (in years) for the original voices and the manipulated versions. Each of the six voices are depicted by a different line colour. The triangle symbol denotes a male voice, and a circle symbol denotes a female voice. Each of the nine points for the different voices represents a version of that voice (from left to right; -20%, -15%, -10%, -5%, 0%, +5%, +10%, +15%, and +20%). 95% confidence intervals are also shown.

7.1.3.2.2 Speech Rate: Part Two

The overall mean accuracy scores were entered into a within-subjects ANOVA. This revealed a significant distractor change, *F*(1, 143) = 1023.85, *p* < .001, , η_g^2 = .87[9]. Voices were estimated as being significantly older when they were manipulated by -20% (*M* = 55.08, *SD* = 6.53) compared to the original voices (*M* = 46.63, *SD* = 5.21), *t*(143) = -25.99, *p* <.001, *d* = 1.43. Voices were also estimated as being significantly older when they were manipulated by -15% (*M* = 52.49, *SD* = 6.28) compared to the original voices (*M* = 46.63, *SD* = 5.21), *t*(143) = -21.98, *p* < .001, *d* = 1.02. Voices were estimated as being significantly older when they were

[9] The author is interested in whether manipulations in speech rate within-speakers can increase errors made in perceptions of speaker age. Therefore, only the simple main effects that are of direct interest are reported here (i.e., comparisons were made between the original voice and the manipulated version of that voice).

manipulated by -10% (*M* = 50.00, *SD* = 6.04) compared to the original voices (*M* = 46.63, *SD* = 5.21), *t*(143) = -16.86, *p* < .001, *d* = 0.60. Voices were also estimated as being significantly older when they were manipulated by -5% (*M* = 48.02, *SD* = 5.72) compared to the original voices (*M* = 46.63, *SD* = 5.21), *t*(143) = -8.78, *p* < .001, *d* = 0.25. In contrast, voices were estimated as being significantly younger when they were manipulated by 20% (*M* = 40.70, *SD* = 5.22) compared to the original voices (*M* = 46.63, *SD* = 5.21), *t*(143) = 24.39, *p* < .001, *d* = 1.14. Voices were also estimated as being significantly younger when they were manipulated by 15% (*M* = 42.15, *SD* = 5.21) compared to the original voices (*M* = 46.63, *SD* = 5.21), *t*(143) = 20.35, *p* < .001, *d* = 0.86. Voices were estimated as being significantly younger when voices they were manipulated by 10% (*M* = 43.60, *SD* = 5.14) compared to the original voices (*M* = 46.63, *SD* = 5.21), *t*(143) = 17.34, *p* < .001, *d* = 0.59. Voices were also estimated as being significantly younger when they were manipulated by 5% (*M* = 45.11, *SD* = 5.26) compared to the original voices (*M* = 46.63, *SD* = 5.21), *t*(143) = 10.53, *p* < .001, *d* = 0.30.

A paired samples *t*-test was also conducted to compare whether there was any difference in age estimates between male and female voices. There was a significant difference in age estimates for male and female voices. Specifically, female voices were estimated as being significantly younger (*M* = 41.16, *SD* = 6.19) compared to male voices (*M* = 53.01, *SD* = 6.00), *t*(143) = 22.66, *p* < .001, *d* = 0.70.

7.1.4 Discussion

The present experiment investigated whether manipulations in F0 or speech rate affect perceptions of speaker age. The results showed that manipulations in both F0 and speech rate changed the perceived age of the speaker. For both male and female voices, increasing F0 and lead to listeners perceiving the voices as sounding younger than the original voices, whereas decreasing F0 lead to listeners perceiving the voices as sounding older than the original voices.

Similarly, for speech rate, increasing the rate of speech lead to listeners perceiving the voices as sounding younger than the original voices, whereas decreasing the rate of speech lead to listeners perceiving the voices as sounding older than the original voices. Therefore, the findings are in line with the original predictions made.

7.1.4.1 Fundamental Frequency (F0)

The findings presented here are consistent with the existing literature for F0 and estimations of speaker age (e.g., Braun & Rietveld, 1995; Hartman & Danhauer, 1976; Horii & Ryan, 1981; Linville & Fisher, 1985; Ptack & Sander, 1966; Shipp et al., 1992). It is important to note that whilst some studies have found no differences in age estimates when F0 is manipulated (Braun & Rietveld, 1995), such discrepancies may be the result of differences in the stimuli and methodology typically employed. For example, several researchers have asked listeners to estimate the age of speakers who differ in chronological age rather than making manipulations in F0 (e.g., Braun & Rietveld, 1995; Ptack & Sander, 1966; Ryan & Burk, 1974). This does not capture within-speaker variations in F0 and how manipulations in F0 might affect age estimations for the same speaker rather than different speakers. Experimental work where the parameter of interest is manipulated constitutes much harder causal evidence for effects of acoustic cues on age estimates (Waller & Eriksson, 2016). This is because any changes in the age estimates can be compared against a control voice (e.g., an original voice) to determine the extent to which manipulations in the cue affect speaker age. Therefore, the present experiment has expanded on the limited work that has been carried out in this domain.

For female voices, the findings follow a similar pattern to the actual differences observed in female voices where F0 continues to drop from childhood to adulthood, and through to older age (refer to Chapter 3, Section 3.1.3.1.1). The findings are also in line with previous research that has found listeners to consistently associate a lower F0 with older age (e.g., Hollien, et al., 2003; Hollien et al., 2008; Smith et al., 2007; Smith & Patterson, 2005; Winkler, 2007). For male voices, whilst the findings are also in line with previous research (e.g., Hollien, et al., 2003; Hollien et al., 2008; Smith et al., 2007; Smith & Patterson, 2005; Winkler, 2007), they do not follow the actual pattern observed for male voices, where F0 decreases from childhood through to adulthood and into middle age, but then rises again into older age (refer to Chapter 3, Section 3.1.3.1.1). The results therefore lend further support to the suggestion that some stereotyping of the vocal characteristics for male speakers may exist (i.e., that decreasing F0 leads to voices being perceived as sounding older regardless of any changes in F0 that actually occur).

Male voices were perceived as sounding significantly older than female voices. This is consistent with the finding in Experiment 4 that, regardless of whether the voice is male or female, voices lower in F0 are perceived as sounding older than voices higher in F0. Indeed, male voices have a lower overall mean F0 compared to female voices, and consequently are perceived as sounding older than female voices. The findings are also in line with previous work that has found listeners to consistently associate a lower F0 with older age (e.g., Hartman & Danhauer, 1976; Waller & Eriksson, 2016).

7.1.4.2 Speech Rate

For both male and female speakers, increasing the rate of speech lead to listeners perceiving the voices as sounding younger than the original voices, whereas decreasing the rate of speech lead to listeners perceiving the voices as sounding older than the original voices. The findings presented here are consistent with previous literature suggesting that listeners are likely to be perceived as sounding older when speech rate is decreased (e.g., Braun & Rietveld, 1995; Hartman & Danhauer, 1976; Horii & Ryan, 1981; Linville & Fisher, 1985; Ptack &

Sander, 1966; Shipp et al., 1992). However, actual changes reported suggest that speech rate increases and reaches its peak value around the mid 40’s (Jacewicz & Fox, 2010; Kowal et al, 1975; Walker et al., 1992), before it begins to get progressively slower into older age (Bruckl & Sendlmeier, 2003; Harnserger, Shrivastav, Brown, Rotham & Hollien, 2006; Linville, 2001; Quene, 2008; Schotz et al., 2006; Verhoeven, De Pauw & Klotts, 2004). The findings in Experiment 4 suggest that even voices that were perceived as younger than the mid-40’s were rated as sounding progressively older as speech rate was decreased. Thus, it is likely that listeners perceive changes in rate occurring at a much younger age than they actually do. This finding lends support to the suggestion that discrepancies may exist between listeners expectations about speakers of different ages and the vocal characteristics that actually exist, and is consistent with previous work that has identified a similar pattern of findings (Hartman & Danhauer, 1976).

Male voices were also perceived as sounding significantly older than female voices when voices were manipulated in speech rate. This finding provides evidence that both F0 and speech rate cues are used to estimate speaker age. Indeed, it is likely that listeners are still using F0 to make estimations about speaker age even though voices were manipulated in speech rate rather than F0. Since male voices have a lower overall mean F0 compared to female voices, male voices are likely to still sound older than female voices when manipulations in speech rate are made. The evidence provided in Experiment 3 (Chapter 6) suggests that F0 is highly indicative of speaker sex, thus it is unlikely that listeners will ignore cues in F0 when male and female voices are heard.

7.1.4.3 Summary Conclusions

The results from the present experiment suggest that manipulations in both F0 and speech rate are likely to affect perceptions of speaker age. For both male and female voices, increasing F0 and speech rate is likely to lead to listeners perceiving the voices as sounding younger, whereas decreasing F0 and speech rate is likely to lead to listeners perceiving the voices as sounding older. However, some discrepancy may exist between listeners expectations about speakers of different ages and the vocal characteristics that actually exist. Therefore, it can be concluded that both F0 and speech rate are important cues for estimating speaker age. The experiments in Chapter's 8 (Experiment 5), 9 (Experiment 6), and 10 (Experiment 7) that follow will move on to consider recognition memory for F0 and speech rate cues of the voice.

CHAPTER 8. VOICE RECOGNITION:

AN EXPLORATION OF THE ACCENTUATION EFFECT

8.1 Experiment 5. An Exploration of the Accentuation Effect: Errors in Memory for Voice Fundamental Frequency (F0) and Speech Rate

8.1.1 Introduction

Chapter 3 (Section 3.2.1) described the principles of the accentuation effect and how memory often reflects category typical representations rather than the specific features of learned items. Whilst there is general agreement within the literature that accentuation effects are real and robust with both social and non-social stimuli, very few researchers have considered accentuation effects in relation to voices. Of the few that do exist, research tends to suggest that listeners are susceptible to distortions in memory for certain properties of the voice, and particularly F0 (Mullenix et al., 2001; Stern et al., 2007). However, given the shortage of studies that have considered accentuation effects in relation to voices, it is difficult for any in-depth conclusions to be drawn.

The present study aimed to investigate the impact of manipulations in F0 or speech rate for a set of unfamiliar synthesised voices in a similar manner to Mullenix et al. (2010), but with a number of important extensions and modifications to the procedure. Mullenix et al. (2010) created a number of versions of a male synthesised target voice; a version that was higher than the original voice and fell within the higher F0 speaking range (which they labelled ‘high F0’), a version that was lower than the original voice and fell within the lower F0 speaking range (labelled ‘low F0’), and the original version of the voice which fell in the moderate F0 speaking range (labelled ‘moderate F0’). Similar manipulations were also applied for the speech rate condition to obtain target voices that were faster in rate (labelled ‘fast rate’), slower in rate (labelled ‘slow rate’), and the original version (labelled ‘moderate rate’). This resulted in six conditions of interest (i.e., high, moderate, and low F0, and fast, moderate, and slow speech rate). Using a two-alternative forced choice (2AFC) voice recognition task, participants were presented with one of the target voices and were then asked to recognise this from a pair of sequentially presented voices. The paired voices included the previously heard target voice and a distractor voice which consisted of a modulated version of the target (which was either higher or lower in F0, or faster or slower in speech rate). In the present experiment, a slightly larger set of synthesised voices (two male, two female) was used, which increases the generalisability of the findings. Second, the target and distractor voices were kept within a F0 and speech rate range that is typical of the human vocal range. This is important given that it is highly unusual to hear voices outside of the typical male and female range in everyday situations. Third, sex of voice and listener sex were included as independent variables in the design. This is important given that research has emphasised sex differences in verbal episodic memory tasks, with females often performing at a higher level than males (Herlitz, Nilsson, & Backman, 1997; Lewin, Wolgers, & Herlitz, 2001; McGivern, Huston, Byrd, King, Siegle, & Reilly, 1997). Others have also reported an own-gender bias (i.e., better recognition performance for voices of an observers own sex) for unfamiliar voices (Roebuck & Wilding, 1993).

Following Mullenix et al. (2010), a 2AFC procedure was used in which listeners were asked to recognise a target voice that had been paired with a modulated version of the voice. There were six conditions of interest (i.e., high, moderate, and low F0, and fast, moderate, and slow speech rate). In keeping with the terminology used by Mullenix et al. (2010), three versions for each target voice were created for both the F0 and speech rate conditions. For the F0 condition, a version was created that was higher than the original voice and fell within the higher F0 speaking range (labelled ‘high F0’), a version that was lower than the original voice and fell within the lower F0 speaking range (labelled ‘low F0’), and the original version of the voice which fell in the moderate F0 speaking range (labelled ‘moderate F0’). Similarly, for the speech rate condition, a version was created that was faster than the original voice (labelled ‘fast rate’), a version that was slower than the original voice (labelled ‘slow rate’), and the original version of the voice (labelled ‘moderate rate’). To obtain the distractor voices, each target voice was further increased and decreased in F0 or speech rate (*for F0:* +/- 5%, +/- 7%, and +/- 10%, *for speech rate:* +/- 10%, +/- 12%, and +/- 20%).

8.1.1.1 Hypotheses

It was expected that for F0, the results would parallel those of Mullenix et al. (2010). It was predicted that there would be a memory bias for high and low F0 target voices but not for moderate F0 target voices. Specifically, it was expected that there would be an increase in the selection of voices higher in F0 when high F0 target voices were presented, and an increase in the selection of voices lower in F0 when low F0 target voices were presented. Although Mullenix et al. (2010) found no memory biases for their speech rate manipulations, consistent with the accentuation effect and the predictions for F0, it was hypothesised that people would be more likely to select distractors that were faster in rate for voices that had a fast speech rate, and to select distractors slower in rate for voices that had a slow speech rate. Furthermore, consistent with much of the existing literature, it was expected that male listeners would make more errors for female voices, whereas female listeners would make more errors for male voices.

8.1.2 Method

8.1.2.1 Design

The participants were allocated to either the F0 condition or the speech rate condition. The F0 condition will subsequently be referred to as Experiment 5a, and the speech rate condition will be referred to as Experiment 5b. For each condition, the experiment employed a 2 x 2 x 3 x 3 x 2 mixed factorial design. The between subjects factor was listener sex (male or female). The within subjects factors were sex of voice (male or female), target type (*for F0:* high, moderate or low*, for speech rate:* fast, moderate or slow), magnitude of distractor change (*for F0:* 5%, 7%, or 10%, *for speech rate:* 10%, 12%, or 20%) and direction of manipulation (*for F0:* increased or decreased in F0, *for speech rate:* increased or decreased in speed). The dependent variable measured was mean percentage of errors made (i.e., percentage of time listeners choose the distractor voice instead of the target voice).

8.1.2.2 Participants

A total of 60 undergraduate students (30 males; 30 females) were recruited from Nottingham Trent University, receiving course credit for their participation. The inclusion criteria for the study required individuals to be between 18 and 30 years of age, have no known hearing deficits, have English as their first language, not undergone any musical training, and had not heard the stimuli presented in the experiment before.

A total of 30 individuals contributed to the F0 condition (Experiment 5a) (15 males; 15 females). The ages of the participants ranged from 18 to 27 years old (M = 21.03 years, SD = 2.09 years). A further 30 individuals contributed to the speech rate condition (Experiment 5b) (15 males; 15 females). The ages of the participants ranged from 18 to 30 years old (M = 21.72 years, SD = 2.62 years).

8.1.2.3 Stimuli and Materials

Of the six original synthesised voices, four were used for experimentation (2 male; 2 female). The two male voices and two female voices with the highest naturalness ratings were chosen for experimentation (refer to Chapter 4, Section 4.3.3). For each of the original synthesised voices, the 10% manipulated versions were used to obtain target voices in the higher and lower F0 range, and the 20% manipulated versions were used to obtain target voices in the faster and slower speech rate range (refer to Chapter 4, Section 4.1.1 and 4.1.2). All four original voice samples fell within the moderate speaking range for both F0 and speech rate, and thus acted as moderate target voices. This resulted in six experimental conditions of interest; low F0, moderate F0 and high F0, and slow speech rate, moderate speech rate, and fast speech rate.

To obtain the distractor speech samples, each target voice was further increased and decreased by 5%, 7%, and 10% for F0, and by 10%, 12%, and 20% for speech rate. The manipulations made were based on the results from previous experimentation (refer to Chapter 4, Section 4.3.1) and allowed the author to determine whether the distractor voices chosen for the experiment were discriminable from the target voice. This resulted in a total of six manipulated versions (i.e., distractor voices) for each target voice sample; three increased in F0 or speech rate, and three decreased in F0 or speech rate (refer to Appendix A1 and A3 for F0 and speech rate values of the voices). All of the voice samples, whether target voices or distractor voices, fell within the typical F0 and speech rate range for normally voiced speech (for F0, the typical adult male will have an F0 between 80-180 Hz, and for an adult female this will be between 165-255 Hz (Titze, 1994), and for speech rate, the typical range for male and female speech is 3.3 to 5.9 syllables/sec (Arnfield, Roach, Setter, Greasley, & Horton, 1995). The distractor stimuli spoke the same phrase as the target stimuli.

All of the speech samples were presented binaurally using Sony dynamic stereo headphones (Model No. MDR-V150). The experiment was run on a Sony Vaio laptop computer (Model No. SVF153B1YM) using PsychoPy version 1.7701 (Peirce, 2007) to control the presentation and collect participant responses.

8.1.2.4 Procedure

The participants were arbitrarily allocated to either the F0 (Experiment 5a) or the speech rate condition (Experiment 5b). Specifically, there were four different voices (two male and two female), each with three target voices (high, moderate, and low F0, or fast, moderate, and slow speech rate). For each target voice, there were 12 trials in total (each of the three target voices were paired with one of the six distractor voices, with each trial being presented twice). As illustrated in Figure 8.1, in each trial, participants were first presented with one of the target speech samples. The text ‘Target Voice’ was visible in the middle of the screen while the target voice was playing. After a one second gap, the participants were presented with sequentially paired voices that included the target voice (present in all trials) and one of the six distractor voices (that was the same voice (i.e., the same speaker) either increased or decreased in F0 or speech rate). There was a one second inter-stimulus interval between presentation of each voice. The text ‘Voice 1’ was visible in the middle of the screen while the first voice was playing, and the text ‘Voice 2’ was visible in the middle of the screen while the second voice was playing. The trials were counterbalanced so that half the time the target voice was presented first, and half the time the target voice was presented second. The order of the trials were randomised across participants using PsychoPy (Pierce, 2007). Following presentation of each trial the participants were asked ‘which voice matched the voice you previously heard, voice one or voice two?’). had to indicate whether the first or the second voice matched the target voice by pressing ‘1’ or ‘2’ on the laptop numerical keyboard. The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life. Upon completion of the experiment, participants were fully debriefed and thanked for their time and participation.

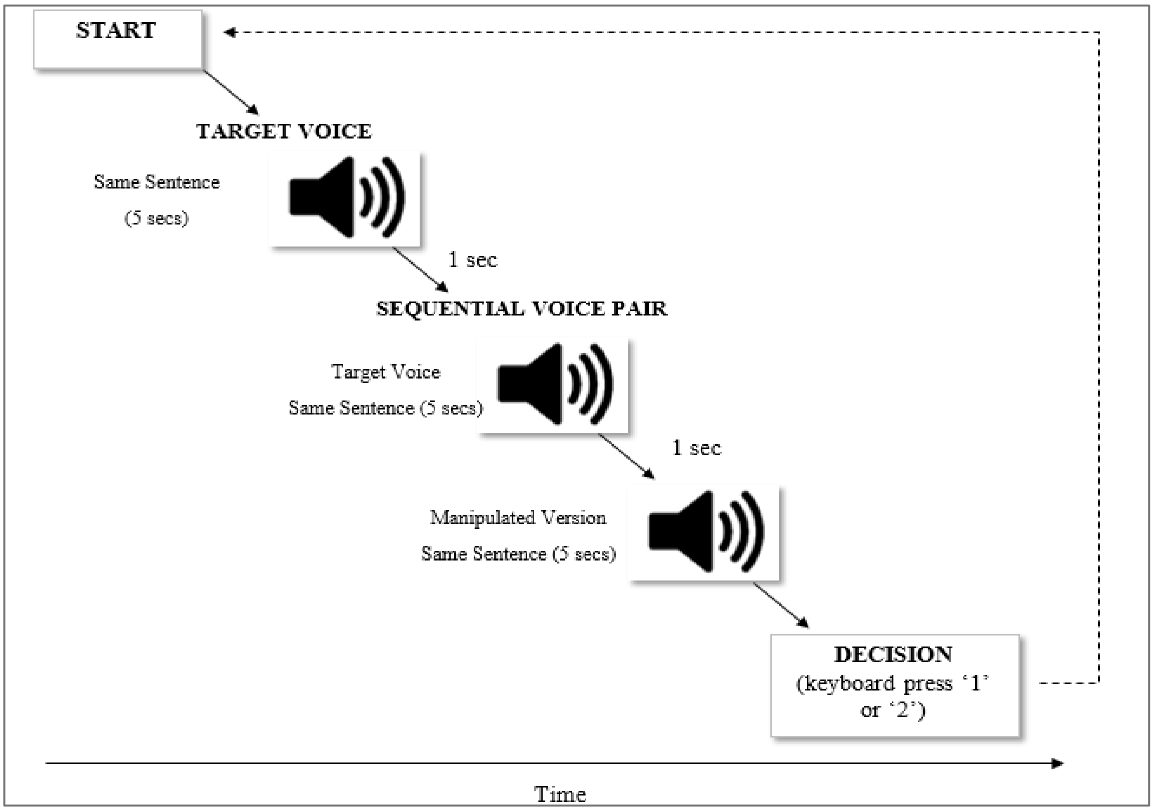


Figure 8.1: An illustration of the procedure in Experiment 5.

8.2.1.5 Analyses

The results were analysed using mixed-group ANOVA, one for the F0 manipulations (Experiment 5a) and one for the speech rate manipulations (Experiment 5b). Owing to the high number of main effects and possible interactions, it was necessary to adjust the *p*-values from the main analysis to account for the familywise error rate. A Hochberg correction was therefore applied to the results of the main ANOVA (Hochberg, 1988). In addition, a Hochberg correction was applied to the simple main effects, which were conducted using pairwise *t*-tests. Furthermore, and for reasons of clarity, only the significant findings of the analyses, or where non-significant findings are directly relevant, are presented here. Full ANOVA tables displaying the degrees of freedom (*df*), F ratios (*F*), effect sizes (generalised eta squared; η_g^2), and adjusted *p* values using the Hochberg correction (*p*) for all the main study variables and associated interactions are provided in Appendix D1 (for F0) and Appendix D2 (for speech rate).

8.1.3 Results

8.1.3.1 Experiment 5a: Fundamental Frequency (F0)

Table 8.1 presents the mean percentage of errors made for each distractor voice, listed separately for the three target conditions (high, moderate and low F0), the sex of the target voice, and listener sex.

The mean matching error scores for each listener were entered into a mixed ANOVA for the between-subjects factor of listener sex (male or female) and the within-subjects factors of sex of voice (male or female), target F0 (high, moderate or low), magnitude of distractor change (5%, 7%, or 10%) and direction of manipulation (increase or decrease in F0). This revealed a significant main effect of direction of manipulation, $F(1, 28) = 94.56$, $p < .03$, $\eta_g^2 = .07$, with significantly more errors being made when distractor voices were higher in F0 ($M = 21.20$, $SD = 7.38$) than when they were lower in F0 ($M = 10.51$, $SD = 5.01$) [10]. There was also a significant main effect of magnitude of distractor change, $F(2, 56) = 50.75$, $p < .03$, $\eta_g^2 = .13$. Significantly more errors were made when distractor voices were manipulated by 5% ($M = 22.64$, $SD = 7.63$) compared to when they were manipulated by 7% ($M = 15.97$, $SD = 7.31$), $t(29) = 4.74$, $p < .001$, $d = 0.37$, and 10% ($M = 8.96$, $SD = 5.66$), $t(29) = 10.10$, $p < .001$, $d = 0.76$ [11]. Significantly more errors were also made when distractor voices were manipulated by 7% ($M = 15.97$, $SD = 7.31$) compared to when they were manipulated by 10% ($M = 8.96$, $SD= 5.66$), $t(29) = 5.63$, $p < .001$, $d = 0.39$. No other main effects were significant or close to significance (adjusted $p > .93$).

[10] Generalised eta-squared statistics (η_g^2) are reported here in order to facilitate comparison between studies with different designs. Generalised eta-square describes the proportion of sample variance accounted for by an effect in an independent design with no manipulated factors (Olejnik & Algina, 2003).

[11] Cohen's *d* values were determined by calculating the mean difference between the two groups, and then dividing the result by the overall pooled standard deviation from all conditions (for F0 = 17.95, for speech rate = 22.98).

Table 8.1: *Mean percentage of errors made by distractor (magnitude of distractor change and direction of distractor change), target F0 (high, moderate, or low), sex of target voice (collapsed across male and female target voices), and sex of listener (male or female).*

	Male Listener						Female Listener					
	Male Target Voice			Female Target Voice			Male Target Voice			Female Target Voice		
	High	Moderate	Low	High	Moderate	Low	High	Moderate	Low	High	Moderate	Low
Distractor												
+10%	**6.67**	**6.67**	**16.67**	**3.33**	**6.67**	**6.67**	***8.33***	**13.33**	**21.67**	**6.67**	**15.00**	**13.33**
	14.84	*11.44*	*22.49*	*8.80*	*14.84*	*11.44*	*15.43*	*20.85*	*20.85*	*14.84*	*18.42*	*12.91*
+7%	**15.00**	**21.67**	**40.00**	**11.67**	**13.33**	**23.33**	**20.00**	**26.67**	**38.33**	**10.00**	**25.00**	**23.33**
	18.42	*20.85*	*28.03*	*18.58*	*12.91*	*22.09*	*23.53*	*22.09*	*24.76*	*18.42*	*29.88*	*14.84*
+5%	**20.00**	**25.00**	**38.33**	**18.33**	**23.33**	**38.33**	**55.00**	**30.00**	**30.00**	**30.00**	**21.67**	**40.00**
	21.55	*25.99*	*20.85*	*19.97*	*17.59*	*26.50*	*28.66*	*28.66*	*19.37*	*21.55*	*16.00*	*18.42*
-5%	**8.33**	**21.67**	**13.33**	**23.33**	**13.33**	**5.00**	**10.00**	**25.00**	**10.00**	**15.00**	**18.33**	**10.00**
	15.43	*18.58*	*18.58*	*22.09*	*16.00*	*10.35*	*15.81*	*23.15*	*15.81*	*12.68*	*17.59*	*18.42*
-7%	**1.67**	**10.00**	**13.33**	**3.33**	**11.67**	**8.33**	**10.00**	**18.33**	**10.00**	**6.67**	**11.67**	**10.00**
	6.46	*18.42*	*12.91*	*8.80*	*16.00*	*15.43*	*18.42*	*17.59*	*22.76*	*11.44*	*22.89*	*15.81*
-10%	**6.67**	**13.33**	**8.33**	**6.67**	**5.00**	**50.00**	**6.67**	**6.67**	**6.67**	**6.67**	**11.67**	**6.67**
	14.84	*18.58*	*18.09*	*14.84*	*10.35*	*14.02*	*11.44*	*14.84*	*14.84*	*14.84*	*16.00*	*11.44*

Note: Means are shown in bold. Standard deviations (SD) are shown in italics.

In addition to the main effects, there was a significant interaction between target F0 and direction of manipulation, $F(2, 56) = 9.27$, $p < .05$, $\eta_g^2 = .03$ [12]. As can be seen in Figure 8.2, the listeners selected higher F0 distractors more often than they selected lower F0 distractors. This effect was strongest for low F0 target voices, with more errors being made when target voices were paired with distractors higher in F0 ($M = 27.50$, $SD = 10.63$) than distractors lower in F0 ($M = 8.89$, $SD = 9.11$), $t(29) = 8.37$, $p < .001$, $d = 1.04$. A similar pattern of findings was apparent for high F0 target voices, with more errors being made when target voices were paired with distractors higher in F0 ($M = 17.08$, $SD = 12.09$) than distractors lower in F0 ($M = 8.75$, $SD = 7.48$), $t(29) = 3.73$, $p < .01$, $d = 0.46$. More errors were also made when moderate F0 target voices were paired with distractors higher in F0 ($M = 19.03$, $SD = 13.07$) than distractors lower in F0 ($M = 13.89$, $SD = 9.21$), $t(29) = 2.40$, $p < .05$, $d = 0.29$.

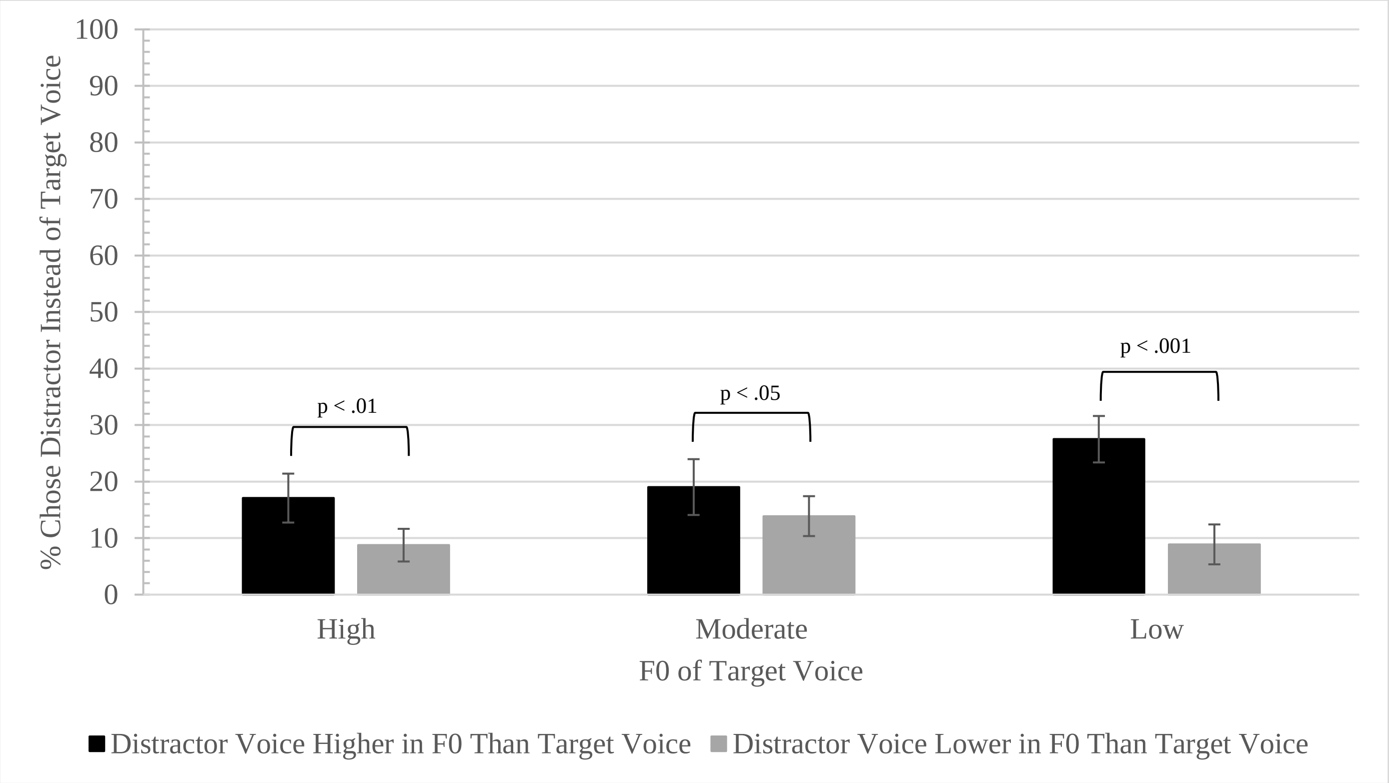


Figure 8.2. Mean percentage of errors made (i.e., chose distractor voice instead of target voice) for the three F0 target voice conditions. 95% confidence intervals are also shown.

[12] The tests of simple main effects that follow are again adjusted using the Hochberg correction. Note that the Hochberg correction is not conditional on a significant *F* ratio in order to protect the Type 1 error. The author corrected for all six possible simple main effects. However, for reasons of brevity only the three simple main effects that are of direct interest are reported here.

There was also a significant interaction between direction of manipulation and magnitude of distractor change, $F(2, 56) = 18.01$, $p < .03$, $\eta_g^2 = .04$ [13]. Figure 8.3 shows that listeners selected distractor voices higher in F0 more often than they selected distractor voices lower in F0 when identifying target voices. This effect was strongest for distractor voices that sounded more similar in F0 to target voices. Specifically, listeners made more errors for distractor voices higher in F0 ($M = 30.83$, $SD = 11.24$) than distractor voices lower in F0 ($M = 14.44$, $SD = 7.24$) when distractor voices were manipulated by 5%, $t(29) = 8.05$, $p < .001$, $d = 0.91$. Listeners also made more errors for distractor voices higher in F0 ($M = 22.36$, $SD = 10.64$) than distractor voices lower in F0 ($M = 9.58$, $SD = 6.58$) when distractor voices were manipulated by 7%, $t(29) = 7.02$, $p < .001$, $d = 0.72$. A similar pattern of findings was also observed for distractor voices that sounded less similar in F0 to target voices (i.e., manipulated by 10%), with more errors being made for distractor voices higher in F0 ($M = 10.42$, $SD = 6.17$) than distractor voices lower in F0 ($M = 7.50$, $SD = 7.37$), $t(29) = 2.13$, $p < .05$, $d = 0.16$.

No other interaction effects were significant or close to significance (adjusted $p > .31$).

[13] The author corrected for all nine possible simple main effects. However, for reasons of brevity only the three simple main effects that are of direct interest are reported here.

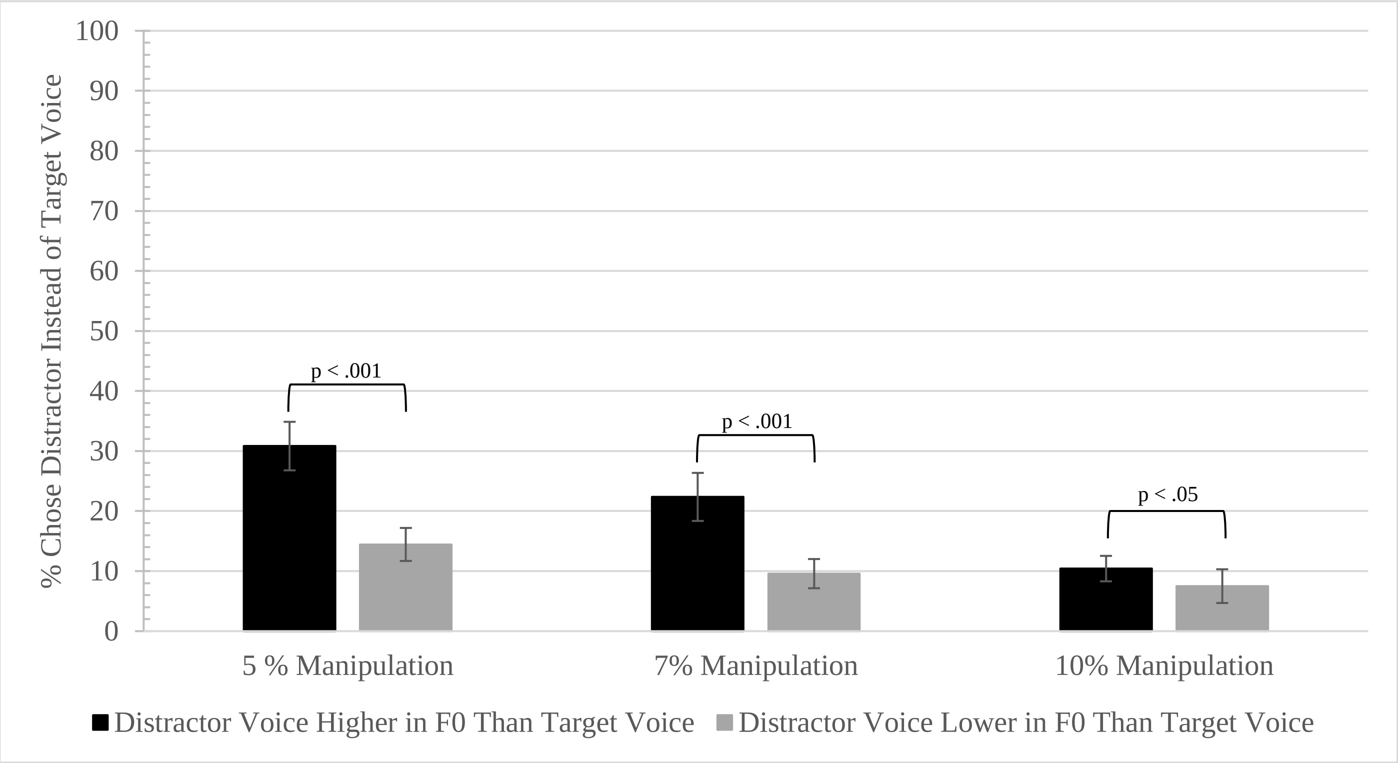


Figure 8.3. Mean percentage of errors made for F0 (i.e., chose distractor voice instead of target voice) for the 5%, 7%, and 10% distractor manipulations. 95% confidence intervals are also shown.

8.1.3.2 Experiment 5b: Speech Rate

Table 8.2 presents the mean percentage of errors made for each distractor voice, listed separately for each of the three target conditions (fast, moderate and slow speech rate), the sex of the target voice, and listener sex.

The matching error scores for each listener were entered into a mixed ANOVA for the between-subjects factor of listener sex (male or female) and the within-subjects factors of sex of voice (male or female), target speech rate (fast, moderate or slow), magnitude of distractor change (10%, 12%, or 20%) and direction of manipulation (increase or decrease in rate). This revealed a significant main effect of direction of manipulation, $F(1, 28) = 12.55$, $p < .05$, η_g^2 =

.02, with significantly more errors being made when the distractor voices were faster in speech rate (*M* = 30.56, *SD* = 7.48) than when they were slower in speech rate (*M* = 25.05, *SD* = 8.06).

There was also a main effect of magnitude of distractor change, *F*(1, 28) = 50.27, *p* < .05, η_g^2 = .10. Significantly more errors were made when distractor voices were manipulated by 10% (*M* = 33.69, *SD* = 8.35) compared to when they were manipulated by 20% (*M* = 18.75, *SD* = 7.95), *t*(29) = 9.90, *p* < .001, *d* = 0.65. Significantly more errors were also made when distractor voices were manipulated by 12% (*M* = 30.97, *SD* = 8.29) compared to when they were manipulated by 20% (*M* = 18.75, *SD* = 7.95), *t*(29) = 9.03, *p* < .001, *d* = 0.53. However, there were no differences in errors made for distractor voices manipulated by 10% (*M* = 33.69, *SD* = 8.35) and 12% (*M* = 30.97, *SD* = 8.29), *t*(29) = 1.52, *p* > .05, *d* = 0.12.

No other main effects were significant or close to significance (adjusted *p* >.96).

Table 8.2. *Mean percentage of errors made by distractor (magnitude of distractor change and direction of distractor change), target speech rate (fast, moderate, or slow), sex of target voice (collapsed across male and female target voices) and sex of listener (male or female).*

	Male Listener						Female Listener					
	Male Target Voice			Female Target Voice			Male Target Voice			Female Target Voice		
	Fast	Moderate	Slow	Fast	Moderate	Slow	Fast	Moderate	Slow	Fast	Moderate	Slow
Distractor												
+20%	**20.00**	**11.67**	**35.00**	**30.17**	**10.00**	**33.33**	**15.00**	**18.33**	**30.00**	**25.00**	**18.33**	**30.00**
	21.55	*20.85*	*29.58*	*28.79*	*18.33*	*24.40*	*18.42*	*14.84*	*25.36*	*32.73*	*24.03*	*25.36*
+12%	**40.00**	**26.67**	**25.00**	**30.00**	**30.00**	**48.33**	**35.00**	**33.33**	**26.67**	**30.00**	**25.00**	**45.00**
	26.39	*22.09*	*16.37*	*27.06*	*28.66*	*22.09*	*18.42*	*26.16*	*17.59*	*27.06*	*18.90*	*19.37*
+10%	**33.33**	**43.33**	**41.67**	**30.17**	**30.00**	**51.67**	**38.33**	**35.00**	**40.00**	**30.17**	**26.67**	**45.00**
	26.16	*33.36*	*18.09*	*28.79*	*25.36*	*25.82*	*24.76*	*28.03*	*29.58*	*28.79*	*17.59*	*23.53*
-10%	**33.33**	**43.33**	**20.00**	**26.67**	**33.33**	**21.67**	**41.67**	**28.33**	**23.33**	**40.00**	**25.00**	**30.00**
	27.82	*29.07*	*21.55*	*22.09*	*26.16*	*22.89*	*34.93*	*26.50*	*25.82*	*22.76*	*18.90*	*19.37*
-12%	**38.33**	**28.33**	**16.67**	**35.00**	**20.00**	**36.67**	**35.00**	**23.33**	**23.33**	**38.33**	**20.00**	**35.00**
	20.85	*28.14*	*20.41*	*24.64*	*23.53*	*22.89*	*29.58*	*25.82*	*14.84*	*26.50*	*25.36*	*22.76*
-20%	**21.67**	**13.33**	**6.67**	**26.67**	**13.33**	**15.00**	**23.33**	**15.00**	**10.00**	**23.33**	**10.00**	**6.67**
	22.89	*12.91*	*14.84*	*24.03*	*20.85*	*18.42*	*17.59*	*22.76*	*20.70*	*19.97*	*15.81*	*11.44*

Note: Means are shown in bold. Standard deviations (SD) are shown in italics.

In addition to the main effects, there was also a significant interaction between target speech rate and direction of manipulation, $F(2, 56) = 15.12$, $p < .05$, $\eta_g^2 = .06$ [14]. Figure 8.4 shows that for slow speech rate target voices, listeners selected distractors faster in rate (M = 37.64, SD = 11.13) more often than they selected distractors slower in rate (M = 20.42, SD = 10.68), $t(29) = 6.34$, $p < .001$, $d = 0.75$. However, there was no difference in the selection of distractors faster in rate (M = 28.35, SD = 14.86) and distractors slower in rate (M = 31.94, SD = 13.33) for fast speech rate target voices, $t(29) = -1.22$, $p >.05$, $d = 0.16$. Furthermore, there was no difference in the selection of distractors faster in rate (M = 25.69, SD = 13.97) and distractors slower in rate (M = 22.78, SD = 12.89) for moderate speech rate target voices, $t(29) = 1.20$, $p > .05$, $d = 0.13$.

No other interaction effects were significant or close to significance (adjusted $p > .43$).

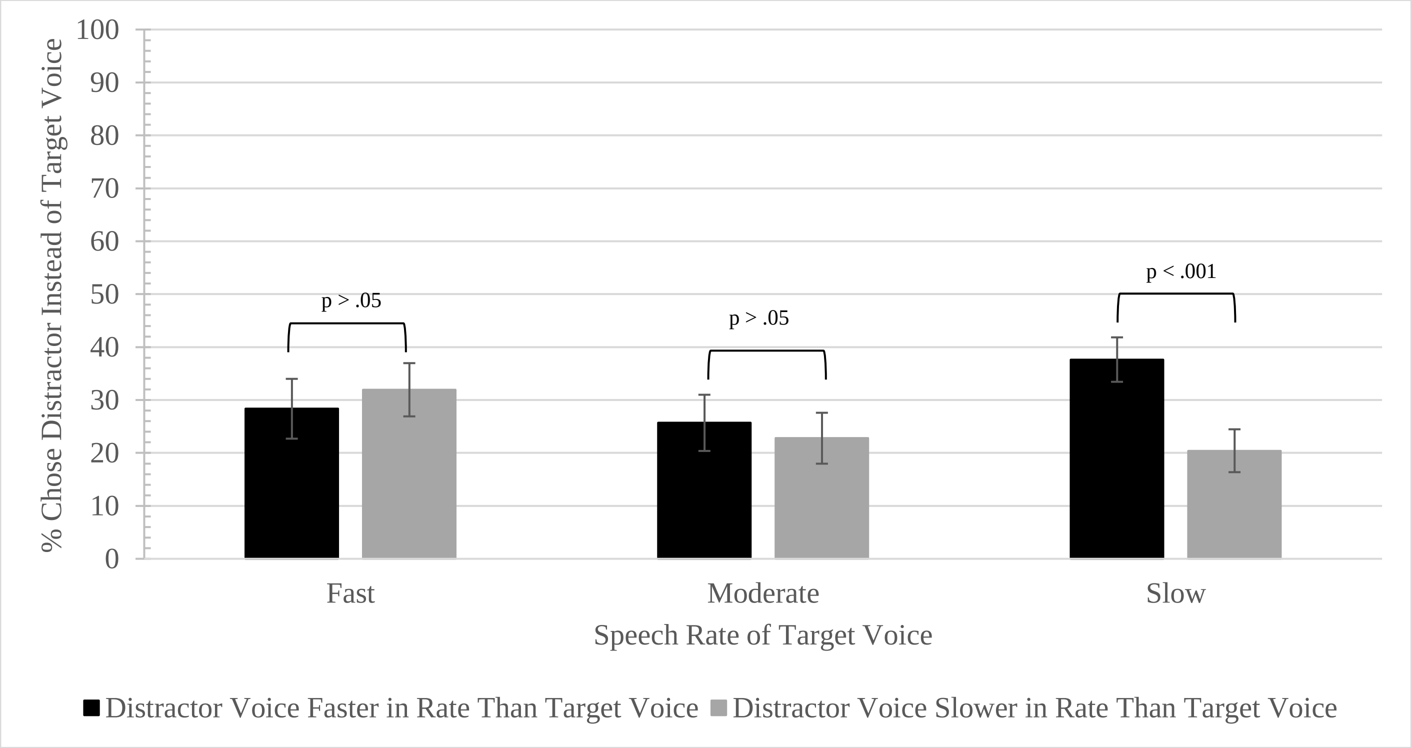


Figure 8.4. Mean percentage of errors made (i.e., chose distractor voice instead of target voice) for the three speech rate target voice conditions. 95% confidence intervals are also shown.

[14] The author corrected for all nine simple main effects. However, for reasons of brevity only the three simple main effects that are of direct interest are reported here.

8.1.4 Discussion

Experiment 5a and 5b investigated the impact of manipulations in F0 or speech rate on immediate target matching performance (selecting a voice from a pair to match a previously heard target voice) for a range of unfamiliar synthesised voices. The findings indicated that there was an increase in the selection of voices higher in F0 when high, moderate, and low F0 target voices were presented. For speech rate, there was an increase in the selection of voices faster in speech rate when slow speech rate target voices were presented. However, no such effect was detected for fast and moderate speech rate target voices. Therefore, in terms of the original hypotheses, there was no evidence for accentuation effects for voice memory. Furthermore, for both the F0 and speech rate conditions, more errors were made identifying target voices when paired with distractor voices manipulated by a smaller magnitude (i.e., 5% for F0, and 10% for speech rate) compared to those manipulated by a greater magnitude (i.e., 10% for F0, and 20% for speech rate). This is perhaps unsurprising given that the results from the pilot study suggest that voices manipulated by a smaller magnitude are harder to distinguish between because they sound more similar to original voices than voices manipulated by a greater magnitude. Thus, more errors are likely to be made identifying target voices when paired with distractor voices manipulated by a smaller magnitude because any differences between the voices are more difficult to detect. There was no effect of either sex of voice or listener sex on errors made identifying target voices.

8.1.4.1 Fundamental Frequency (F0)

The results presented here do offer some support to those identified by Mullenix et al. (2010) in that errors in memory are likely to occur for voice F0. However, the finding of an increase in the selection of voices higher in F0 is difficult to explain using the accentuation effect. It is unlikely that this outcome is an anomaly in the data set given that the findings are reasonably consistent across all target voices. They are also unlikely to be the result of order effects because the voices presented in the voice pair were counterbalanced across participants. Given that synthesised voices were used for experimentation, some of the acoustic properties of the stimuli could explain the observed pattern of findings. However, this is unlikely given that the voices used were rated as sounding natural (refer to Chapter 4, Section 4.3.4), formant frequencies changed freely, formant transitions were smooth, and there were no intonational irregularities or prosodic mismatches across words. This alleviates concerns that something uncontrolled and artificial about the stimuli were driving the findings. Rather, the extensions and modifications made to the study procedure may explain the difference in results. First, the target and distractor voices within a F0 range that is typical in the population (i.e., between 80-180 Hz, and for an adult female this will be between 165-255 Hz (Titze, 1994). In contrast, the manipulations made by Mullenix et al. (2010) fell considerably outside of this range. Second, a set of four synthesised voices were used, whereas Mullenix et al. (2010) used only a single voice. Therefore, it is quite possible that the findings identified by Mullenix et al. (2010) were due to the peculiarity of the stimuli (i.e., an unusually high or low F0) used in the experiment. Using a more representative and generalizable set of voices, as in the present study (i.e., a slightly larger set of synthesised voices, with manipulations in F0 or speech rate kept within a range that is typical in the population for English speakers), the accentuation bias is no longer found. The data reported here suggest little or no accentuation bias for the memory of voice F0.

It is not entirely clear why listeners make more errors recognising voice F0 when paired with distractor voices higher in F0 compared to when they are paired with distractor voices lower in F0. However, it is quite possible that the listeners had difficulty discriminating between the frequencies of some of the voice pairs in the experiment. Indeed, research has identified that it is more difficult to discriminate between voices of higher frequencies compared to voices of lower frequencies (Moore, 1995). In the present study, listeners may have made fewer errors identifying target voices when paired with distractor voices lower in F0 because they were more efficient at detecting the changes in frequency than when distractor voices were higher in frequency. This interpretation would account for why there was no effect of listener sex on errors made identifying target voices, because there is no reason to believe that the perceptual capabilities of the listener would differ substantially between male and female listeners. It would also explain why there was no difference in errors made for male and female target voices. Although female voices are higher in F0 than male voices, the findings are based upon a listener's ability to detect any *differences* in the frequencies of the voices in the voice pair, and this is independent of the frequency of the target voice itself.

It is also likely that listeners made more errors identifying target voices when paired with distractor voices higher in F0 compared to when they are paired with distractor voices lower in F0 because they resemble voices that are typically heard in the general population. Inflection refers to the frequency patterns in a person's speech, where the voice rises and falls, either upwards or downwards in frequency (Fairbanks, 1940). Research has shown that all types of inflections are greater in upward inflection than they are in downward inflection (e.g., Benjamin, 1981; Fairbanks & Pronovost, 1939). Furthermore, researchers have shown that when people are asked to choose a method of disguise, they are more likely to raise the frequency of their voice rather than lowering it (e.g., Mathur, Choudhary, & Vyas, 2016; Masthoff, 1996). Such evidence suggests that people are more likely to increase, rather than decrease, the frequency of their voice when they speak. Thus, the listeners in the present study may be selecting distractor voices higher in F0 more often than distractor voices lower in F0 because they are more familiar with these types of utterances and it sounds like a more plausible version of the target voice (i.e., an inflected version of the target voice).

The finding that listeners were more likely to select distractor voices higher in F0 compared to distractor voices lower in F0 was particularly prevalent for the low F0 target voice condition. This bias may have arisen because voices higher in F0 are perceived as less threatening than voices lower in F0. Research has shown that both male and female voices lowered in F0 are perceived as more dominant, threatening, and aggressive than the same voices raised in F0 (e.g., Borkowska & Pawlowski, 2011; Fraccaro, O'Connor, Re, Jones, DeBruine, & Feinberg, 2012; Jones, Feinberg, DeBruine, Little, & Vukovic, 2010; Morton, 1994; Ohala, 1984; Puts, Gaulin, & Verdonili, 2006). Furthermore, evidence tends to suggest that people will often exhibit avoidance type behaviour when exposed to aversive stimuli (Corr, 2013). Assuming that in the present study, the voices lower in F0 would be rated as sounding more dominant and threatening than the voices higher in F0, listeners may have selected the higher voice of the pair because it sounded less dominant and less threatening. This would explain why an increase in the selection of higher F0 distractors was particularly prevalent for the low F0 target voice condition; because the voices were decreased in F0 sufficiently for the higher F0 voices in the pair to be perceived as less threatening to the listener. It would also account for why there was no effect of either sex of voice or listener sex; perceptions of dominance have been found to be equivalent for both male and female voices and male and female listeners (Jones et al., 2010). Further work would be required to confirm or disconfirm this explanation to the findings.

Another possibility that also deserves equal consideration for why the selection of distractor voices higher in F0 compared to distractor voices lower in F0 was particularly prevalent for the low F0 target voice condition, is that English voices lower in F0 for both males and females tend to co-occur with covariations in voice quality (e.g., Aberton, Howard, & Fourcin, 1989). A bias towards selecting the higher F0 distractor voices could reflect the unnaturalness of the voices lowered in F0 without a concomitant change in voice quality.

Whilst the voices were rated as sounding natural, this issue might still remain even if naturally sounding voices were modified to have a lower F0.

Finally, it is worth noting that the naturalness ratings for the voices with higher F0 manipulations tended to yield slightly higher naturalness rating scores than those with lower manipulations (refer to Chapter 4, Section 4.3.4). One possible interpretation of this is that the listeners preferred the more natural sounding voices (i.e., the higher F0 manipulations) and were thus, more likely to select them. Unfortunately, because the naturalness ratings came from a different population to those in the 2AFC tasks reported here, it was not appropriate to formally test this possibility. The authors tuition, given that the voices were generally perceived to be natural sounding across the board, and any differences observed between the voices were relatively small, are unlikely to have impacted upon the matching tasks. Thus, whilst it is a possibility that naturalness may have an effect, the question is unable to resolved here.

8.1.4.2 Speech Rate

For speech rate, listeners selected voices faster in speech rate when slow speech rate target voices were presented. Thus, the findings presented here are difficult to explain with reference to the accentuation effect. Given this, it is possible that the findings could be accounted for by the listener’s level of familiarity of the voice heard. In natural speech, a person speaking more slowly is likely to be more hesitant, making more silent pauses or filled pauses (e.g., *um, er*). In the present study, decreasing speech rate did affect the rate of continuous production but did not lead to increased pauses of any kind. It is therefore unlikely that the speech samples used were an entirely natural rendition of slower speech, at least of a type that listeners most typically hear. It is possible that at the lower margins of the speech rate manipulated samples (i.e., the slowest samples), but not elsewhere, the participants may have selected a faster voice in the pair because it sounded more realistic.

Faster speaking voices might also sound more favourable when compared with slower speaking voices in the slow speech rate pairings. Indeed, research suggests that speech rates can influence a listener’s perceptions of a speaker’s personality and social skills. For example, faster speaking styles have been shown to be rated more favourably (Stewart & Ryan, 1982), and viewed as more competent and socially attractive than voices spoken at a slower rate (Street, Brady, & Putman, 1983). Slower speaking styles have also been identified as sounding weaker, less truthful, and less empathetic than voices spoken at a faster rate (Apple, Streeter, & Krauss, 1979). It is possible that listeners were more likely to select a faster voice in the pair because they preferred the sound of the voice. However, such selections may have been made only for the slow speech rate condition because these voices were slowed sufficiently for the faster rate voices in the pair to be rated more favourably, and thus selected by the listener. The above explanations would also account for why there was no effect of either sex of voice or listener sex on errors made identifying a target voice, as there is no reason to suggest that the level of familiarity or preference for faster voices would differ between male and female voices, or for male and female listeners.

8.1.4.3 Summary Conclusions

The results from the present experiment suggest that listeners are susceptible to distortions in memory for F0 more so than they are for speech rate. However, the data reported here cannot be accounted for in terms of the accentuation effect. Therefore, it is doubtful that listeners rely solely on the self-generated categorical information about the voice at the time of encoding to aid recognition of the voice at a later stage. The present experiment has thus contributed to our understanding of the mechanisms important for accurate voice recognition, and such work may prove as a useful conceptual tool in determining the properties of voice that are more or less affected by intra-individual variation. Experiment 6 (Chapter 9) will move on to consider whether listeners become increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition when the inter-stimulus interval is increased. Given that a more robust pattern of errors was identified in Experiment 5 for manipulations in F0 than for manipulations in speech rate, it was decided that speech rate would be omitted from the experiments that are to follow.

CHAPTER 9. VOICE RECOGNITION:

INCREASING THE INTER-STIMULUS INTERVAL

9.1 Experiment 6. An Exploration of the Accentuation Effect: Errors in Memory for Voice Fundamental Frequency (F0) with an Increased Inter-Stimulus Interval

9.1.1 Introduction

Experiment 5 (Chapter 8) investigated the impact of manipulations in F0 or speech rate for a set of unfamiliar synthesised voices in a similar manner to Mullenix et al. (2010). The study set out to determine whether the accentuation bias can account for any errors in recognition performance that occur. The results showed that there was no evidence for accentuation effects in voice memory for either F0 or speech rate. Instead, for F0, there was an increase in the selection of voices higher in F0 when both high and low F0 target voices were presented. For speech rate, there was an increase in the selection of voices faster in speech rate when slow speech rate target voices were presented.

Chapter 3 (Section 3.2.2.1) discussed the time course of the echoic memory store and the possibility that performance on a memory task for auditory stimuli could be dependent on this. The more precise the mental representation of the auditory stimulus, the more accurate the memory for that stimulus is likely to be. We might expect people to be more accurate in a task that uses shorter inter-stimulus intervals between presentation of auditory stimuli than on one where the interval is longer in duration (Ivry & Spencer, 2004; Mauk & Buonomano, 2004; Buhusi & Meck, 2005; van Wassenhove, 2009). What’s more, any bias affecting performance may also be dependent on this time course. Experiment 5 used a 1-second inter-stimulus interval between presentation of the target voice and the sequential voice pair (i.e., the previously heard target voice and a manipulated version of the target voice). The task may have been relatively easy for listeners because the acoustic trace of the voice was likely to still be strong. However, as the inter-stimulus interval increases the task is likely to become more difficult, and listeners may become increasingly reliant on category based information stored in memory to aid recognition. This is because the acoustic trace for the previously heard target voice is likely to be weaker when the inter-stimulus interval between presentations of the voices is increased. We might therefore expect the findings to be more like those evidenced by Mullenix et al. (2010), with listeners selecting distractor voices higher in F0 for high F0 target voices, and selecting distractor voices lower in F0 for low F0 target voices, because they are affected by the accentuation effect.

Therefore, Experiment 6 aimed to determine whether listeners become increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition when the inter-interval stimulus is increased. The impact of manipulations in F0 were investigated in a similar manner to Experiment 5a (Chapter 8), with the addition of a 5-second inter-stimulus interval between presentation of the target voice and the sequential voice pair. Given that a more robust pattern of errors was identified for manipulations in F0 than for manipulations in speech rate, it was decided that speech rate would be omitted from the experiment. This finding has also been supported by the existing literature, suggesting that F0 is more likely to be the variable of interest (e.g., Mullenix et al., 2001; Stern et al., 2007). Therefore, there were only three conditions of interest (i.e. high, moderate, and low F0). As in Experiment 5 (Chapter 8), Experiment 6 used a 2AFC procedure in which listeners were asked to recognise a target voice that had been paired with a manipulated version of the target voice. The stimuli used were identical to those used in the F0 condition (Experiment 5a).

9.1.1.1 Hypotheses

It was anticipated that the inclusion of a five-second inter-stimulus interval would lead to listeners becoming increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition of the voice in the matching task. Therefore, consistent with Mullenix et al. (2010), it was predicted that there would be a memory bias for high and low F0 target voices but not for moderate F0 target voices. Specifically, it was expected that there would be an increase in the selection of voices higher in F0 when high F0 target voices were presented, and an increase in the selection of voices lower in F0 when low F0 target voices were presented. In light of the findings in Experiment 5 (Chapter 8), it was unlikely that there would be an effect of either sex of voice or listener sex on errors made. Nevertheless, these have been included in the analysis for completeness.

9.1.2 Method

9.1.2.1 Design

The experiment employed a 2 x 2 x 3 x 2 x 2 mixed factorial design. The between-subjects factor was listener sex (male or female). The within-subjects factors were sex of voice (male or female), target F0 (high, moderate, or low), magnitude of distractor change (5% or 10%)[15] and direction of manipulations (increased or decreased in F0). The dependant variable measured was mean percentage of errors made (i.e., percentage of time listeners chose the distractor voice instead of the target voice).

9.1.2.2 Participants

A total of 30 undergraduate students (15 males; 15 females) were recruited from Nottingham Trent University and received course credit for their participation. The age of the

[15] There was no difference in errors made for the 5% versus the 7% distractors in Experiment 5a (Chapter 8). Therefore, for ease of interpretation, 7% distractors were not included in this experiment.

participants ranged from 18 to 26 years old (*M* = 21.53 years, *SD* = 2.98 years). The inclusion criteria for the study required individual to be between 18 to 30 years of age, have no known hearing deficits, have English as their first language, not undergone any musical training, and had not heard the stimuli presented in the experiment before.

9.1.2.3 Stimuli and Materials

The stimuli and materials were identical to those use in the F0 condition in Experiment 5a (Chapter 8) (refer to Appendix A1 for further details of the voices).

9.1.2.4 Procedure

The procedure was identical to that in Experiment 5a (Chapter 8) with the exception of there being only 96 trials in total (8 trials for each target voice, with each trial being presented twice). This was because of the focus on the F0 condition only. There was also the addition of a five-second inter-stimulus interval between presentation of each voice in the voice pair. There was a one-second inter-stimulus interval between presentation of each voice in the voice pair. The voices were presented at the same loudness for all participants. This was at a level that was typical of a conversation you would hear in everyday life. Figure 9.1 provides and illustration of the procedure used in Experiment 6.

9.1.2.5 Analyses

The results were analysed in an identical manner to those in Experiment 5 (Chapter 8). Only the significant findings of the analyses, or where non-significant findings are directly relevant, are presented here. Full ANOVA tables displaying degrees of freedom (*df*), F ratios (*F*), effect sizes, η_g^2, and adjusted *p* values using the Hochberg correction (*p*) for all the main study variables and associated interactions are provided in Appendix E.

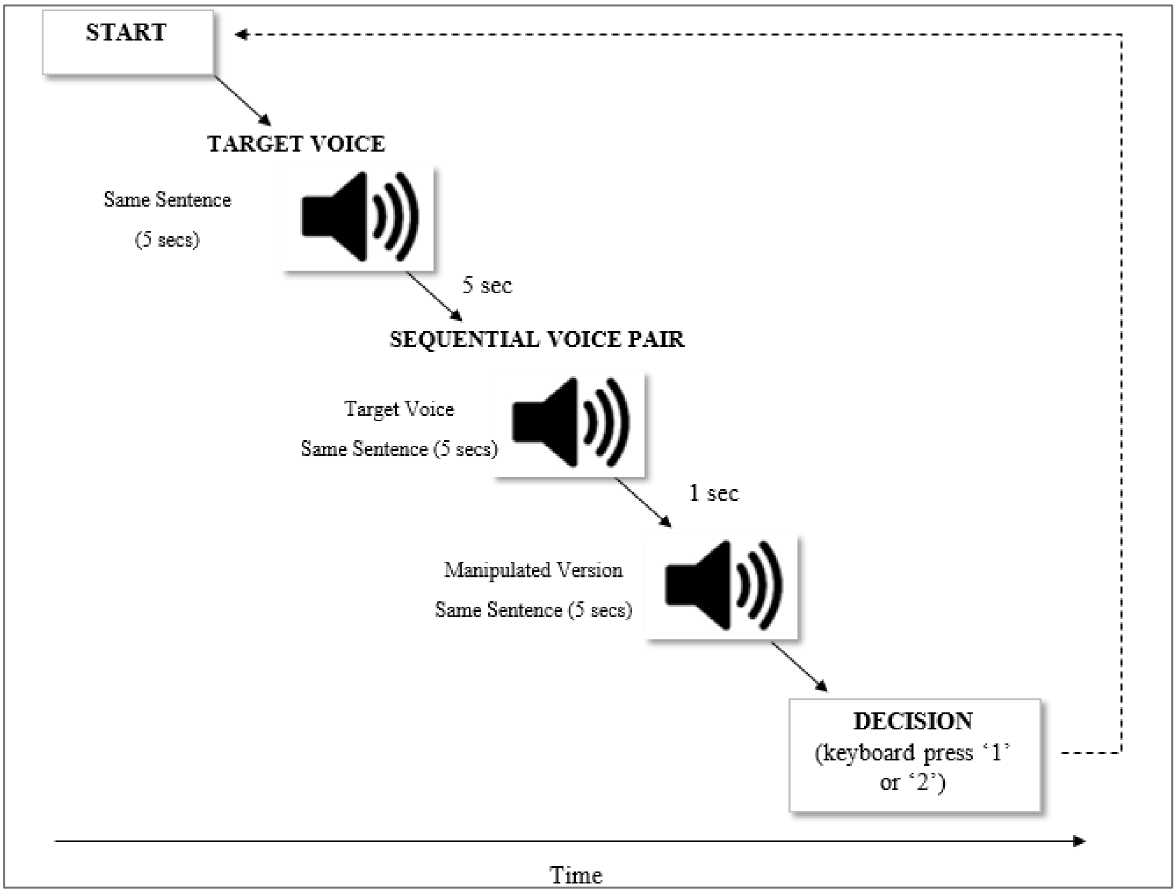


Figure 9.1: An illustration of the procedure in Experiment 6.

9.1.3 Results

9.1.3.1 Fundamental Frequency (F0)

Table 9.1 presents the mean percentage of errors made for each distractor, listed separately for the three target conditions (high, moderate, and low F0), sex of target voice, and listener sex.

Table 9.1: *Mean percentage of errors made by distractor (magnitude of distractor change and direction of distractor change), target F0 (high, moderate, or low), sex of target voice (collapsed across male and female target voices) and sex of listener (male or female).*

	Male Listener						Female Listener					
	Male Target Voice			Female Target Voice			Male Target Voice			Female Target Voice		
	High	Moderate	Low	High	Moderate	Low	High	Moderate	Low	High	Moderate	Low
Distractor												
+10%	**13.33**	**13.33**	**18.33**	**8.33**	**13.33**	**30.00**	**18.33**	**13.33**	**18.33**	**15.00**	**16.67**	**8.33**
	18.58	16.00	*24.03*	*12.20*	*18.58*	*27.06*	*22.09*	*18.58*	*19.97*	*24.64*	*18.09*	*18.09*
+5%	**26.67**	**31.67**	**30.00**	**10.00**	**26.67**	**21.67**	**23.33**	**35.00**	**43.33**	**20.00**	**28.33**	**43.33**
	22.09	*22.09*	*19.37*	*12.68*	*19.97*	*28.14*	*30.57*	*26.39*	*24.03*	*33.00*	*29.68*	*32.00*
-5%	**15.00**	**15.00**	**6.67**	**3.33**	**10.00**	**3.33**	**16.67**	**18.33**	**18.33**	**20.00**	**20.00**	**16.67**
	15.81	*18.42*	*14.84*	*8.80*	*12.68*	*8.80*	*22.49*	*17.59*	*19.97*	*25.36*	*25.36*	*27.82*
-10%	**20.00**	**20.00**	**8.33**	**8.33**	**3.33**	**8.33**	**16.67**	**13.33**	**15.00**	**11.67**	**11.67**	**5.00**
	19.37	*23.53*	*12.20*	*15.40*	*8.80*	*12.20*	*26.16*	*18.58*	*24.60*	*16.00*	*16.00*	*10.35*

Note: Means are shown in bold. Standard deviations (SD) are shown in italics.

The mean matching error scores were entered in a mixed ANOVA for the between-subjects factor of listener sex (male or female) and the within-subjects factors of sex of voice (male or female), target F0 (high, moderate or low), magnitude of distractor change (5% or 10%) and direction of manipulation (increase or decrease in F0). This revealed a significant main effect of direction of manipulation, $F(1, 28) = 27.90$, $p < .05$, $\eta_g^2 = .06$, with significantly more errors being made when distractor voices were higher in F0 ($M = 21.67$, $SD = 12.70$) than when they were lower in F0 ($M = 12.71$, $SD = 11.69$). No other main effects were significant or close to significance (adjusted $p > .056$).

In addition to the main effects, there was a significant interaction between target F0 and direction of manipulation, $F(2, 56) = 11.80$, $p < .05$, $\eta_g^2 = .02$ [16]. As can be seen in Figure 9.2, the listeners selected distractor voices higher in F0 more often than they selected distractor voices lower in F0 when identifying target voices. This pattern was stronger for low F0 target voices, with more errors being made when the target voices were paired with distractors higher in F0 ($M = 27.92$, $SD = 11.46$) than distractors lower in F0 ($M = 10.21$, $SD = 12.97$), $t(29) = 4.64$, $p < .01$, $d = .38$ [17]. A similar pattern of findings was also apparent for moderate F0 target voices, with more errors being made when target voices were paired with distractors higher in F0 ($M = 22.29$, $SD = 13.70$) than distractors lower in F0 ($M = 13.96$, $SD = 11.80$), $t(29) = 3.61$, $p < .01$, $d = .41$. However, this was not the case for high F0 target voices, where there was no difference in the errors made when target voices were paired with distractors higher in F0 (*M*

[16] The tests of simple main effects that follow are again adjusted using the Hochberg correction. Note that the Hochberg correction is not conditional on a significant *F* ratio in order to protect the Type 1 error. We corrected for all six possible simple main effects. However, for reasons of brevity only the three simple main effects that are of direct interest are reported here.

[17] Cohen's *d* values were determined by calculating the mean difference between the two groups, and then dividing the result by the overall pooled standard deviation from all conditions (for F0 = 20.38).

= 16.04, *SD* = 15.46) compared to distractors lower in F0 (*M* = 13.96, *SD* = 14.18), *t*(29) = .89, *p* > .05, *d* = .10.

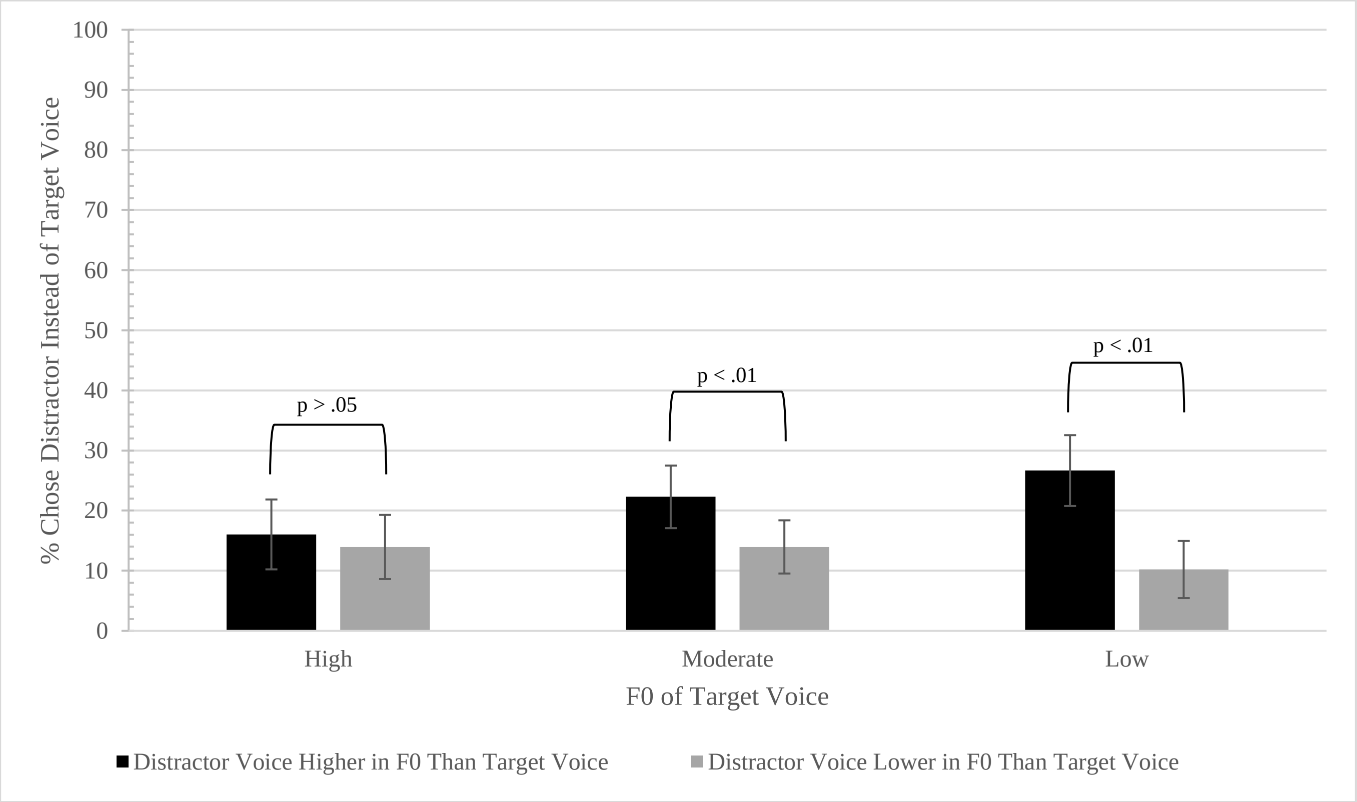


Figure 9.2: Mean percentage of errors made (i.e. chose distractor voice instead of target voice) for the three F0 target voice conditions. 95% confidence intervals are also shown.

There was also a significant interaction between direction of manipulation and magnitude of distractor change, *F*(1, 28) = 18.90, *p* < .05, η_g^2 = .02. Figure 9.3 shows that the listeners selected distractor voices higher in F0 more often than they selected distractors lower in F0 when identifying target voices. However, this effect was only significant for distractor voices that sounded more similar to the target voices (i.e. manipulated by 5%), with more errors being made for distractor voices higher in F0 (*M* = 28.33, *SD* = 17.25) than distractor voices lower in F0 (*M* = 13.61, *SD* = 14.18), *t*(29) = 6.88, *p* < .01, *d* = .72. The listeners also selected distractor voices higher in F0 than target voices more often when the distractor voices sounded similar to target voices (i.e., manipulated by 5%) (*M* = 28.33, *SD* = 17.25) than when the distractor voices sounded less similar to target voices (i.e., manipulated by 10%) (*M* = 15.00, *SD* = 12.88), *t*(29) = 4.35, p < .01, *d* = .64. No other comparisons were significant (*p* > .05).

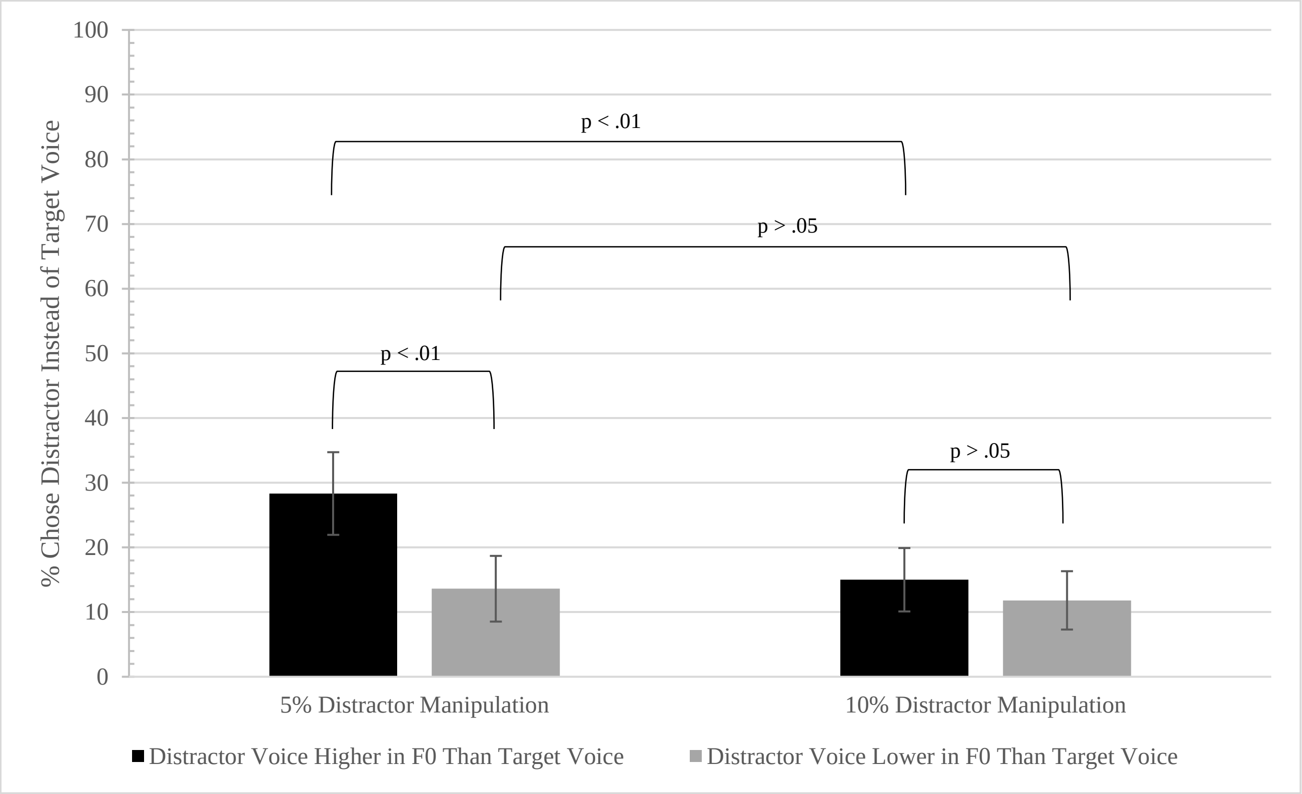


Figure 9.3: Mean percentage of errors made (i.e. chose distractor voice instead of target voice) for the 5% and 10% distractor manipulations. 95% confidence intervals are also shown.

No other interaction effects were significant or close to significance (adjusted *p* > .96).

9.1.4 Discussion

The pattern of findings in Experiment 6 are similar to those observed in Experiment 5a, where there was a 1-second interval between hearing the target voice and being presented with the sequential voice pair. The results from Experiment 6 showed that there was an increase in the selection of voices higher in F0 when both moderate and low F0 target voices were presented. In contrast, no such effect was found for high F0 target voices. Therefore, in terms of the original hypotheses, there was no evidence for accentuation effects for the memory of voice F0 when the interval between hearing the target voice and being asked to recognise this from a voice pair was increased to five seconds. Thus, listeners are no more likely to rely on self-generated categorical information about the voice at the time of encoding to aid recognition when the inter-stimulus interval is increased. There was no effect of either sex of voice or listener sex on errors made identifying target voices.

The finding of an increase in the selection of voices higher in F0 than the target voice appears to be reasonably consistent across Experiments 5a and 6, suggesting that this is not an anomaly in the data set and is a robust outcome. Furthermore, it lends support to several conclusions drawn in Experiment 5a. First, that the accentuation bias is no longer found when a more representative and generalisable set of voices are used, and second, that listeners have difficulty discriminating between voices of higher frequencies. Listeners may have made fewer errors identifying target voices when paired with distractor voices lower in F0 because they were more efficient at detecting the changes in frequency than when distractor voices were higher in frequency. Again, this interpretation would account for why there was no effect of listener sex or sex of voice on errors made identifying target voices. Finally, the findings in Experiment 6 demonstrated that the tendency for listeners to select voices higher in F0 in the voice pair was strongest for the low F0 target voice condition. This lends additional support to the assumption made in Experiment 5a that voices higher in F0 are chosen by the listeners because they are perceived as less threatening or dominant sounding to the listener.

There was no difference in errors made when high F0 target voices were paired with distractor voices that were higher or lower in F0 than target voices. However, the pattern identified in Experiment 6 is similar to that seen in Experiment 5a (Chapter 8). Given this, the most likely explanation for any difference in the findings across the two experiments was that the effect in Experiment 6 happened to not be significant this time.

More errors were made identifying target voices when paired with distractor voices manipulated by a smaller magnitude (i.e., 5%) compared to those manipulated by a greater magnitude (i.e., 10%). This is perhaps unsurprising given that the results from Experiment 1a suggest that voices manipulated by a smaller magnitude (i.e., 5%) are harder to differentiate between, as they sound more similar to the original voices than voices manipulated by a greater magnitude (i.e., 10%). Furthermore, the findings in Experiment 2 (Chapter 5) suggest that voices manipulated by a greater magnitude (i.e., 10%) are more likely to sound like a different speaker compared to voices manipulated by a smaller magnitude (i.e., 5%).

9.1.4.1 Summary Conclusions

The results from the present experiment suggest that listeners are susceptible to distortions in memory for F0. However, the data reported here cannot be accounted for in terms of the accentuation effect. Therefore, it is doubtful that listeners become increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition of the voice when the inter-stimulus interval is increased to five seconds. The present experiment has thus contributed to our understanding of the mechanisms important for accurate voice recognition, and such work may prove as a useful conceptual tool in determining the properties of voice that are more or less affected by intra-individual variation. Experiment 7 (Chapter 10) will move on to consider whether listeners become increasingly reliant on self-generated information about the voice at the time of encoding to aid recognition when a different sentence is spoken in the sequential voice pair to the one previously heard. In addition to this experiment, an additional experiment will be conducted to determine whether the amount of errors made overall differed between the three recognition memory experiments (Experiment 5a in Chapter 8, Experiment 6 in Chapter 9, and Experiment 7) when manipulations in F0 were made.

CHAPTER 10. VOICE RECOGNITION: CHANGING THE SPOKEN MESSAGE

10.1. Experiment 7. An Exploration of the Accentuation Effect: Errors in Memory for Voice Fundamental Frequency (F0) with a Different Sentence

10.1.1 Introduction

The experiments carried out so far have demonstrated little evidence for accentuation effects in voice memory for F0. Both Experiment 5a (Chapter 8) and Experiment 6 (Chapter 9) showed a very similar pattern of findings; listeners tend to select voices higher in F0 compared to voices lower in F0 for the three target voice conditions. In Experiment 5a there was an increase in the selection of voices higher in F0 when high, moderate, and low F0 target voices are presented. In Experiment 6, there was an increase in the selection of voices higher in F0 when both moderate and low F0 target voices are presented. This effect appears to be particularly apparent for the low F0 target voice condition in both Experiment 5a (Chapter 8) and Experiment 6 (Chapter 9). Experiment 5b (Chapter 8) showed that for speech rate, there was an increase in the selection of voices faster in rate when slow speech rate target voices were presented.

Both Experiment 5a (Chapter 8) and Experiment 6 (Chapter 9) used the same sentence for the target voice and the voices in the sequential voice pair (i.e., the previously heard target voice and a manipulated version of the target voice). Retrieval of a voice from memory may be somewhat easier if the sentence spoke at the matching phase is the same as the previously heard target sentence. Recognition of the voice might be accomplished on the basis of a simple familiarity judgement, and without any knowledge of the voice per se, if the same sentence is used (refer to Chapter 3, Section 3.2.2.2). By repeating the same sentence, listeners can use patterns identified in the spoken sentence to determine whether the voice heard matches the mental representation of the voice stored in memory. However, using a different sentence at the matching phase to the one previously heard is likely to make the task harder by ensuring that recognition of the voice is not dependent on an exact match of the content in the sentence (Glisky, Rubin, & Davidson, 2001). Therefore, any bias affecting performance may also be dependent on the sentence used. When a different sentence is used to the one that was previously heard, it is more difficult for listeners to make judgements about the voice because they cannot rely on the linguistic content of the sentence. Accordingly, listeners are likely to become increasingly reliant on non-linguistic properties (i.e., F0) of the voice. Thus, category typical representations stored in memory may be used more when making decisions about the voice, resulting in a bias of the characteristics properties of the voice. Consequently, there may be an increase in accentuation errors when matching a voice to a previously heard target voice.

Therefore, Experiment 7 aimed to determine whether listeners become increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition when a different sentence is spoken to the one previously heard. The impact of manipulations in F0 were investigated in a similar manner to Experiment 5a (Chapter 8). Again, given that a more robust pattern of errors was identified for manipulations in F0 than for manipulations in speech rate, it was decided that speech rate would be omitted from the experiment. Therefore, there were three conditions of interest (i.e., high, moderate, and low F0). As in Experiment 5a (Chapter 8), Experiment 7 used a 2AFC procedure in which listeners were asked to recognise a target voice that had been paired with a manipulated version of the target voice. However, this time as part of the sequential voice pair, the participants were presented with voices that spoke a non-identical speech phrase to the previously heard target voice. The additional versions of the target speech samples were created by making the same manipulations used to obtain the previous target voices in Experiment 5a (Chapter 8). This resulted in a version that was higher than the original voice (labelled high F0), a version that was lower than the original voice (labelled low F0), and the original version of the voice which fell in the moderate F0 speaking range (labelled moderate F0). To obtain the distractor voices, each of the additional target voices were further increased and decreased in F0.

10.1.1.1 Hypotheses

It was anticipated that the inclusion of a different sentence spoken at the testing phase would result in listeners becoming increasingly reliant on categorical information self-generated about the voice at the time of encoding to aid recognition of the voice at a later stage. Therefore, consistent with Mullenix et al. (2010), it was predicted that there would be a memory bias for high and low F0 target voices but not for moderate F0 target voices. Specifically, it was expected that there would be an increase in the selection of voices higher in F0 when high F0 target voices were presented, and an increase in the selection of voices lower in F0 when low F0 target voices were presented. In light of the findings in Experiment 5 (Chapter 8) and Experiment 6 (Chapter 7), it was unlikely that there would be an effect of either sex of voice or listener sex on errors made. Nevertheless, these have been included in the analysis for completeness.

10.1.2 Method

10.1.2.1 Design

The experiment employed a 2 x 2 x 3 x 2 x 2 mixed factorial design. The between-subjects factor was listener sex (male or female). The within-subjects factors were sex of voice (male or female), target F0 (high, moderate or low), magnitude of distractor change (5% or

10%)[18] and direction of manipulation (increased or decreased in F0). The dependent variable measured was mean percentage of errors made (i.e., percentage of time listeners choose the distractor voice instead of the target voice).

10.1.2.2 Participants

A total of 30 undergraduate students (15 males; 15 females) were recruited from Nottingham Trent University and received course credit for their participation. The ages of the participants ranged from 19 to 29 years old (M = 23.52 years, SD = 2.17 years). The inclusion criteria for the study required individuals to be between 18-30 years of age, have no known hearing deficits, have English as their first language, not undergone any musical training, and had not heard the stimuli presented in the experiment before.

10.1.2.3 Stimuli and Materials

The materials and stimuli were identical to those used for the F0 condition in Experiment 5a (Chapter 8) and Experiment 6 (Chapter 9) with the exception of a different speech phrase being used in the sequential voice pair. The additional speech samples were created by typing the chosen phrase in Natural Reader. For each of the four original voices, the phrase *"Living costs have more than tripled, and gas has gone down one third*", was used to create the target speech samples that listeners would hear as part of the sequential voice pair. The length of the speech samples was 5 seconds in length. This matched the duration of the sentences used in Experiment 5a (Chapter 8) and Experiment 6 (Chapter 9). In keeping with the labels used by Mullenix et al. (2010), these voices contributed to the moderate F0 target voice condition. To obtain target voices that fell within the higher and lower F0 range (and would contribute to the high and low F0 target voice conditions) each of the four target voices

[18] Again, there was no difference in errors made for the 5% versus the 7% distractors in Experiment 5a (Chapter 8). Therefore, for ease of interpretation, the 7% distractors were not included in this experiment.

were manipulated by +10% (i.e., increased in F0) and –10% (i.e., decreased in F0) using Audacity® software. To obtain the distractor speech samples, each target voice was further increased and decreased by 5% and 10%. This resulted in a total of four manipulated versions (i.e., distractor voices) for each target voice sample; two increased in F0 and two decreased in F0 (refer to Appendix A3 for further details of the voices).

10.1.2.4 Procedure

The procedure was identical to that in Experiment 5 (Chapter 8) with the exception of there being only 96 trials in total (8 trials for each target voice, with each trial being presented twice). This was because the focus was on the F0 condition only. However, as part of the sequential voice pair, the participants were presented with voices that spoke a non-identical speech phrase to the previously heard target voice. There was a one-second inter-stimulus interval between presentation of each voice. The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life. Figure 10.1 provides an illustration of the procedure used in Experiment 7.

10.1.2.5 Analyses

The results were analysed in an identical manner to those in Experiment 5 (Chapter 8) and Experiment 6 (Chapter 9). Only the significant findings of the analyses, or where non-significant findings are directly relevant, are presented. Full ANOVA tables displaying the degrees of freedom (*df*), F ratios (*F*), effect sizes (generalised eta squared; η_g^2), and adjusted *p* values using the Hochberg correction (*p*) for all the main study variables and associated interactions are provided in Appendix F.

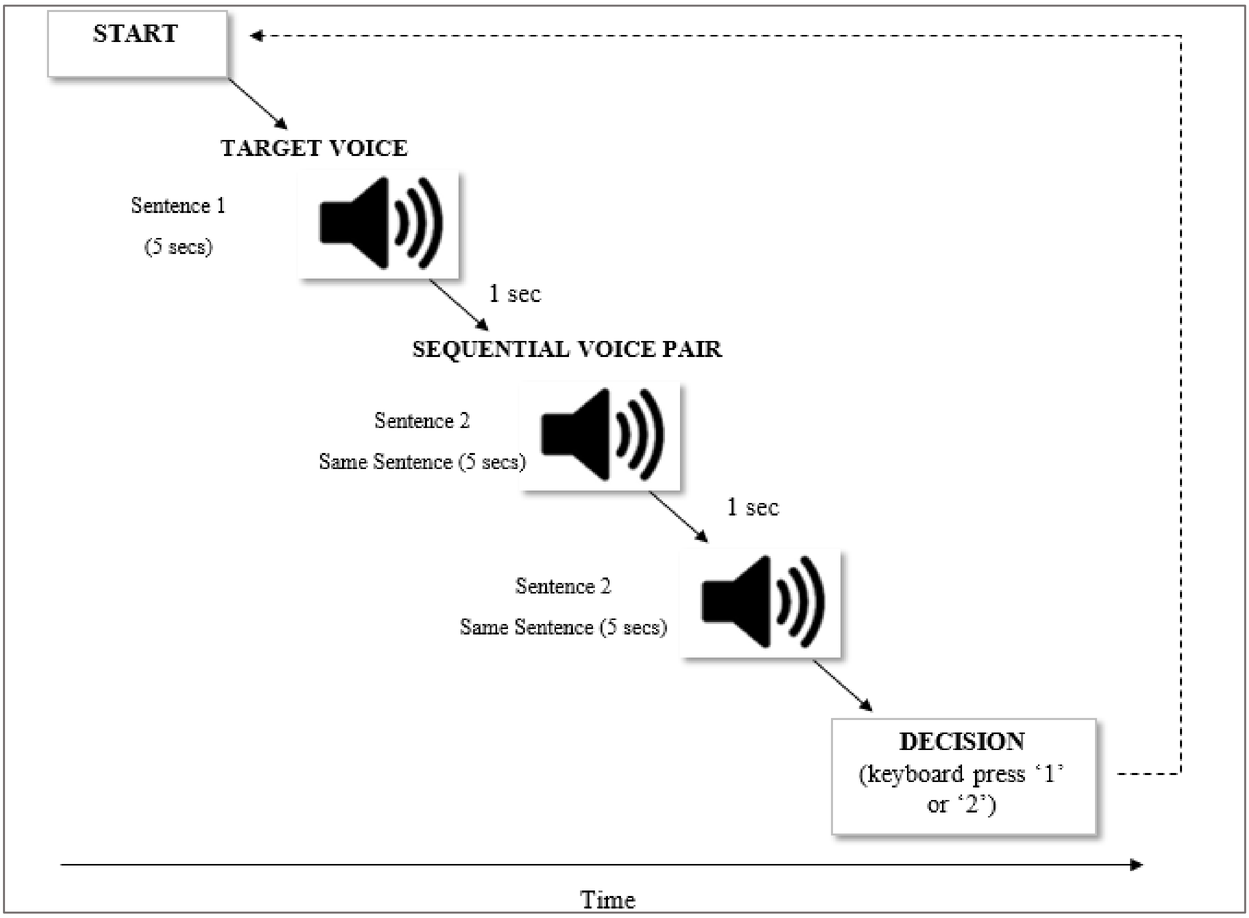


Figure 10.1: An illustration of the procedure in Experiment 7.

10.1.3 Results

10.1.3.1 Fundamental Frequency (F0)

Table 10.1 presents the mean percentage of errors made for each distractor, listed separately for the three target conditions (high, moderate, and low F0), sex of target voice, and listener sex.

Table 10.1: *Mean percentage of errors made by distractor (magnitude of distractor change and direction of distractor change), target F0 (high, moderate, or low), sex of target voice (collapsed across male and female target voices) and sex of listener (male or female).*

	Male Listener						Female Listener					
	Male Target Voice			Female Target Voice			Male Target Voice			Female Target Voice		
	High	Moderate	Low	High	Moderate	Low	High	Moderate	Low	High	Moderate	Low
Distractor												
+10%	**10.00**	**11.67**	**16.67**	**20.00**	**20.00**	**35.00**	**11.67**	**16.67**	**25.00**	**18.33**	**36.67**	**45.00**
	18.42	*18.58*	*24.40*	*16.90*	*21.55*	*20.70*	*16.00*	*15.43*	*23.15*	*24.03*	*24.76*	*30.18*
+5%	**23.33**	**26.67**	**40.00**	**25.00**	**30.00**	**48.33**	**21.67**	**30.00**	**40.00**	**23.33**	**48.33**	**63.33**
	17.59	*17.59*	*22.76*	*21.13*	*25.36*	*30.57*	*20.85*	*23.53*	*28.03*	*19.97*	*14.84*	*26.50*
-5%	**35.00**	**18.33**	**10.00**	**23.33**	**18.33**	**5.00**	**26.67**	**15.00**	**16.67**	**15.00**	**16.67**	**13.33**
	24.60	*24.03*	*12.68*	*25.82*	*17.59*	*10.35*	*24.03*	*15.81*	*18.09*	*22.76*	*20.41*	*16.00*
-10%	**20.00**	**8.33**	**8.33**	**15.00**	**8.33**	**3.33**	**21.67**	**16.67**	**13.33**	**23.33**	**16.67**	**5.00**
	21.55	*15.43*	*20.41*	*15.81*	*12.20*	*8.80*	*26.50*	*20.41*	*18.58*	*17.59*	*26.16*	*14.02*

Note: Means are shown in bold. Standard deviations (SD) are shown in italics.

The mean matching error scores were entered in a mixed ANOVA for the between-subjects factor of listener sex (male or female) and the within subjects factors of sex of voice (male or female), target F0 (high, moderate or low), magnitude of distractor change (5% or 10%) and direction of manipulation (increase or decrease in F0). As can be seen in Figure 10.2, this revealed a significant main effect of magnitude of distractor change, $F(1, 28) = 49.60$, $p < .05$, $\eta_g^2 = .04$, with significantly more errors being made when distractor voices sounded more similar to target voices (i.e., manipulated by 5%) ($M = 26.39$, $SD = 7.80$) than when they sounded less similar to target voices (i.e., manipulated by 10%) ($M = 17.78$, $SD = 8.49$).

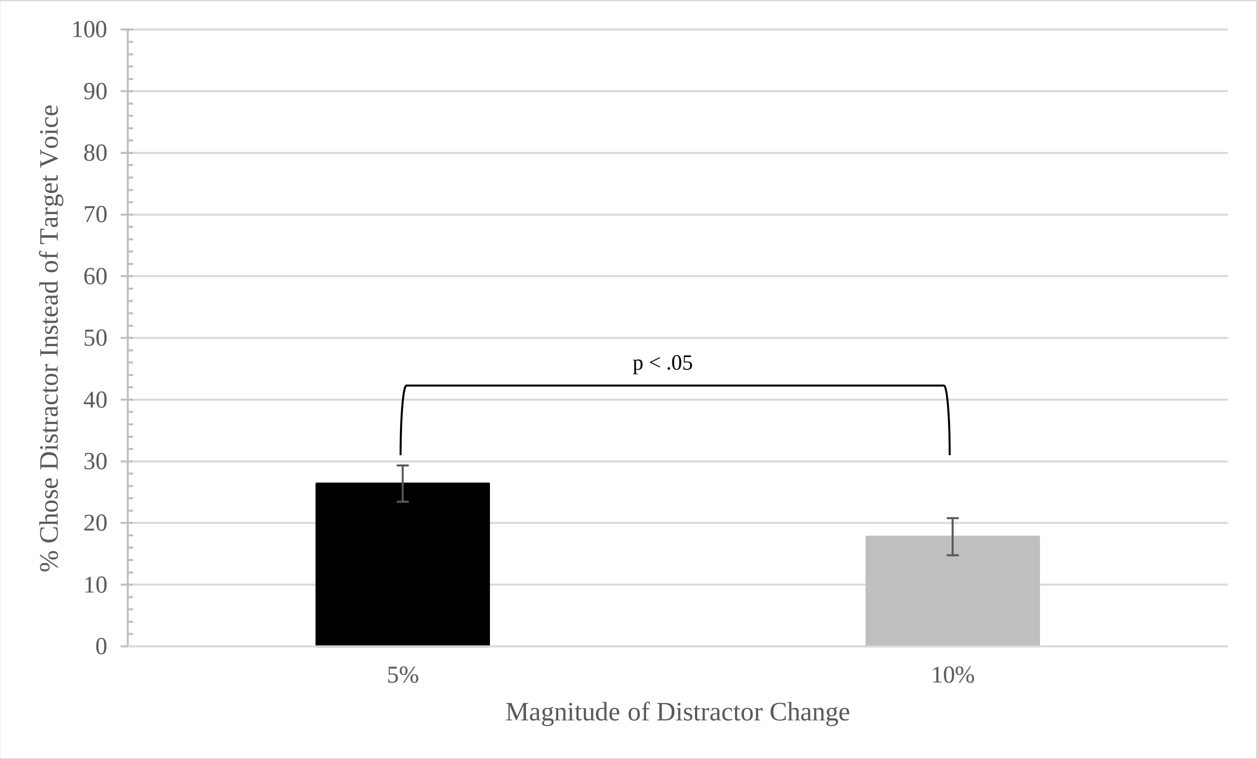


Figure 10.2: Mean percentage of errors made (i.e., chose distractor voice instead of target voice) for the 5% and 10% distractor manipulations. 95% confidence intervals are also shown.

There was also a significant main effect of direction of manipulation, $F(1, 28) = 40.84$, $p < .05$, $\eta_g^2 = .08$, with significantly more errors being made when distractor voices were higher in F0 ($M = 28.61$, $SD = 9.52$) than when they were lower in F0 ($M = 15.56$, $SD = 9.00$). No other main effects were significant or close to significance (adjusted $p > .93$).

In addition to the main effects, there was a significant interaction between target F0 and direction of manipulation, $F(2, 56) = 26.54$, $p < .05$, $\eta_g^2 = .09$ [19]. As can be seen in Figure 10.3, the listeners selected distractor voices higher in F0 more often than they selected distractor voices lower in F0 when identifying target voices. This effect was strongest for low F0 target voices, with more errors being made when target voices were paired with distractors higher in F0 ($M = 39.17$, $SD = 16.97$) than distractors lower in F0 ($M = 9.38$, $SD = 8.65$), $t(29) = 8.80$, $p < .01$, $d = 1.45$ [20]. A similar pattern of findings was also apparent for moderate F0 target voices, with more errors being made when target voices were paired with distractors higher in F0 ($M = 27.50$, $SD = 13.69$) than distractors lower in F0 ($M = 14.79$, $SD = 10.95$), $t(29) = 4.21$, $p < .01$, $d = 0.62$. However, this was not the case for high F0 target voices, where there was no difference in the errors made when target voices were paired with distractors higher in F0 ($M = 19.17$, $SD = 8.52$) compared to distractors lower in F0 ($M = 22.50$, $SD = 15.54$), $t(29) = -.96$, $p > .05$, $d = 0.16$.

[19] The tests of simple main effects that follow are again adjusted using the Hochberg correction. Note that the Hochberg correction is not conditional on a significant *F* ratio in order to protect the Type 1 error. We corrected for all six possible simple main effects. However, for reasons of brevity only the three simple main effects that are of direct interest are reported here.

[20] Cohen's *d* values were determined by calculating the mean difference between the two groups, and then dividing the result by the overall pooled standard deviation from all conditions (for F0 = 20.50).

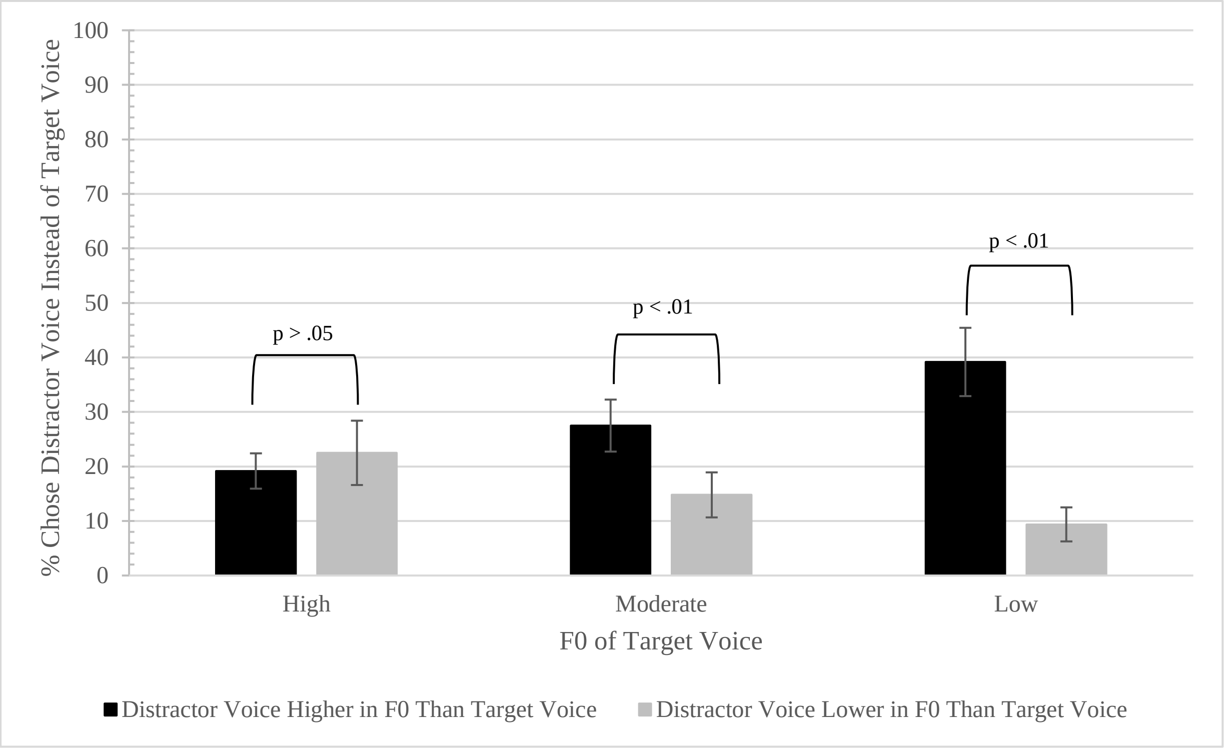


Figure 10.3: Mean percentage of errors made (i.e., chose distractor voice instead of target voice) for the three F0 target voice conditions. 95% confidence intervals are also shown.

There was also a significant interaction between direction of manipulation and sex of voice, $F(1, 28) = 21.34$, $p < .05$, $\eta_g^2 = .03$. Figure 10.4 shows that the listeners selected distractor voices higher in F0 ($M = 34.44$, $SD = 12.66$) more often than they selected distractor voices lower in F0 ($M = 13.61$, $SD = 10.14$) when matching female target voices, $t(29) = 7.47$, $p < .01$, $d = 1.02$. However, this was not the case for male target voices, where there was no difference in the errors made when target voices were paired with distractors higher in F0 (M = 22.78, SD = 11.92) compared to distractors lower in F0 ($M = 17.50$, $SD = 10.96$), $t(29) = 2.13$, $p > .05$, $d = 0.26$. The listeners also selected distractor voices higher in F0 more often than they selected distractor voices lower in F0 when matching female target voices ($M = 34.44$, $SD = 12.66$) compared to male target voices ($M = 22.78$, $SD = 11.92$, $t(29) = -4.11$, $p < .01$, $d = 0.57$. However, this was not the case for distractor voices lower in F0, where there was no difference in the errors made when matching female target voices (*M* = 13.61, *SD* = 10.14) and male target voices (*M* = 17.50, *SD* = 10.96), *t*(29) = 1.93, *p* > .05, *d* = 0.19.

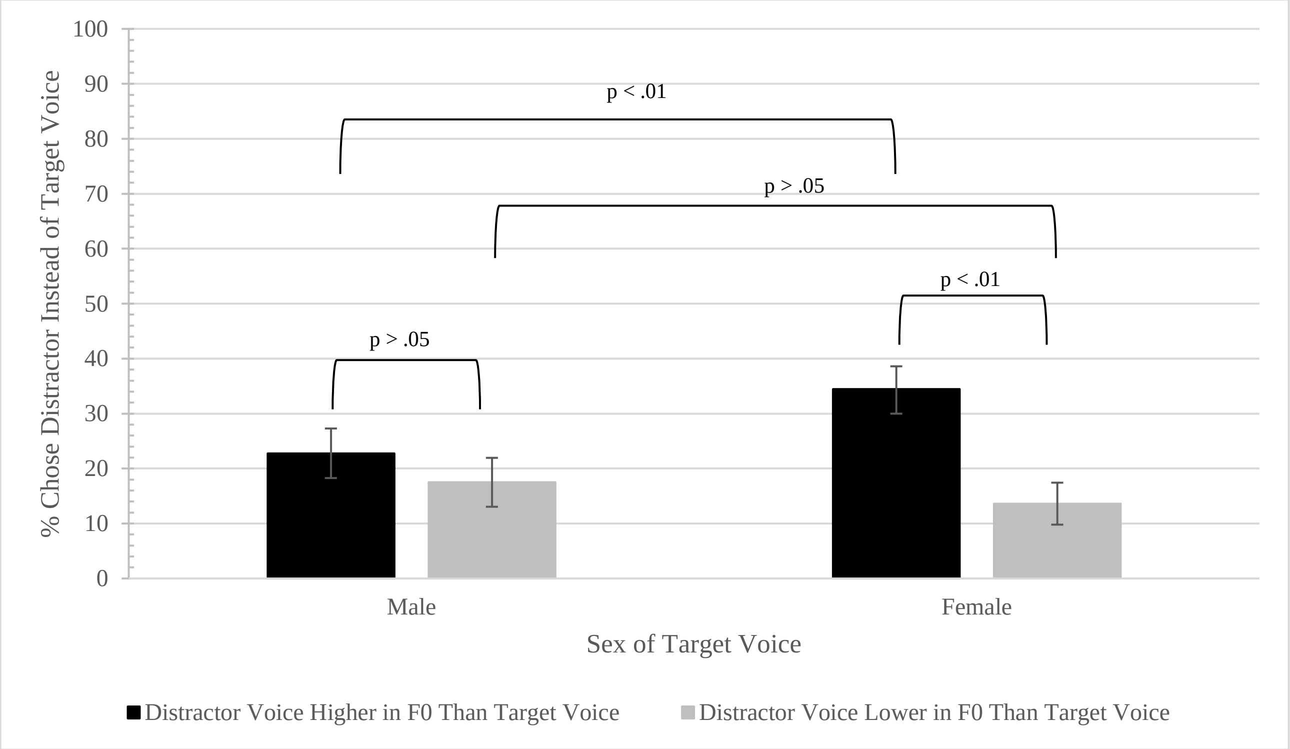


Figure 10.4: Mean percentage of errors made (i.e. chose distractor voice instead of target voice) for male and female voices. 95% confidence intervals are also shown.

No other interaction effects were significant or close to significance (*p* > .40).

10.1.4 Discussion

The pattern of findings in Experiment 7 are similar to those observed in both Experiment 5a and 6, showing that there was an increase in the selection of distractor voices higher in F0 compared to distractor voices lower in F0 when both moderate and low F0 target voices were presented. In contrast, no such effect was found for high F0 target voices. Therefore, in terms of the original hypotheses, there was no evidence for accentuation effects for the memory of voice F0 when a different sentence is spoken to the one previously heard. Thus, listeners are no more likely to rely on self-generated categorical information about the voice at the time of encoding to aid recognition when a different sentence is spoken to the one previously heard.

The findings of an increase in the selection of voices higher in F0 than the target voice appears to be reasonably consistent across Experiments 5a, 6, and 7, lending further support to several conclusions previously drawn. First, that listeners may have made fewer errors identifying target voices when paired with distractor voices lower in F0 compared to distractor voices higher in F0 because they were more efficient at detecting the changes in frequency than when the distractor voices were higher in frequency. Second, that the listeners may be selecting distractor voices higher in F0 more often than distractor voices lower in F0 because they are more familiar with these types of utterances and it sounds like a more plausible version of the target voice (i.e., an inflected version of the target voice). And third, that the tendency for listeners to select voices higher in F0 compared to voices lower in F0 was strongest for the low F0 target voice condition, because voice higher in F0 are perceived as less threatening to the listener.

One important difference in the findings of Experiment 7 compared to those in Experiments 5a and 6, was that there was an increase in the selection of distractor voices higher in F0 compared to distractor voices lower in F0 for female target voices. In contrast, no difference was found for male target voices. One possible explanation for this finding is that higher F0 female voices may have sounded more like those voices that are typically heard. Indeed, it has been reported that females use upward inflections when they speak more than twice as often as males do (Guy, Horvath, Vonwiller, Daisley, & Rogers, 1986; Hoffman, 2013). Therefore, listeners may be selecting distractor voices higher in F0 compared to distractor voices lower in F0 for female voices because they are familiar with this style of utterance and it sounds like a more plausible version of the target voice (i.e., an inflected version of the target voice). This finding may have only occurred in Experiment 7 because listeners are forced to rely on the acoustic properties of the voice (in this case, F0) rather than content when the sentence is different to the one previously heard. Conversely, both Experiment’s 5a and 6, listeners are able to use elements of the spoken sentence *and* F0 to help aid the matching process.

Another explanation for this finding is that male voices were actually increased in F0 less than female target voices. A relative percentage change (rather than an absolute percentage change) was used to manipulate the voices in F0. Thus, a percentage change in F0 for male voices was smaller than the same percentage change for female voices because male voices have a lower overall mean F0. It is possible that fewer errors are made for male voices than for female voices because the manipulated versions were more similar in F0 to the target voice. Because listeners are having to rely more on the acoustic properties of the voice (in this case, F0) than on the content of the sentence, they perform better in the matching task for male voices than they do for female voices.

More errors were made identifying target voices when paired with distractor voices manipulated by a smaller magnitude (i.e., 5%) compared to those manipulated by a greater magnitude (i.e., 10%). This is perhaps unsurprising given that the results from Experiment 1a suggest that voices manipulated by a smaller magnitude (i.e., 5%) are harder to differentiate between, as they sound more similar to the original voices than voices manipulated by a greater magnitude (i.e., 10%). Furthermore, the findings in Experiment 2 (Chapter 5) suggest that voices manipulated by a greater magnitude (i.e., 10%) are more likely to sound like a different speaker compared to voices manipulated by a smaller magnitude (i.e., 5%).

10.1.4.1 Summary Conclusions

The results from the present experiment suggest that listeners are susceptible to distortions in memory for F0. However, the data reported here cannot be accounted for using the accentuation effect. Therefore, it is doubtful that listeners become increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition when a different sentence is used to the one previously heard. However, it appears that listeners may find it more difficult to match female voices when the content of the sentence is changed. Experiment 8 will move on to determine whether the number of errors made overall the three recognition memory experiments (Experiment 5a in Chapter 8, Experiment 6 in Chapter 9, and Experiment 7) were different when manipulations in F0 were made.

10.2. Experiment 8. A Comparison of Recognition Errors across Experiments 5a, 6, and 7

10.2.1 Introduction and Aims

The aim of Experiment 8 was to determine whether the amount of matching errors made overall were different for the three recognition memory experiments (Experiment 5a, Experiment 6, and Experiment 7) when manipulations in F0 were made. Experiment 8 explored whether accurate matching decisions rely on high quality representations temporarily stored in echoic memory (Experiment 5a, Chapter 8) and whether the ability to make accurate decisions diminishes as the inter-stimulus interval increases (Experiment 6, Chapter 9). Experiment 8 also explored whether listeners are more accurate at matching voices when the content of the sentences in the sequential voice pair is the same as the sentence spoken by the target voice (Experiment 5a, Chapter 8) compared to when the content of the sentence in the sequential voice pair is different to the sentence spoken by the target voice (Experiment 7, Chapter 10).

10.2.1.1 Hypotheses

In line with the findings from previous experimentation, it was predicted that there would be no difference in matching errors overall when a 1-second inter-stimulus interval was used (Experiment 5a, Chapter 8) and when a 5-second inter-stimulus interval was used (Experiment 6, Chapter 9). It was also predicted that listeners would make fewer matching errors overall when the content of the sentence in the sequential voice pair was the same as the sentence spoken by the target voice (Experiment 5a, Chapter 8) compared to when the content of the sentence in the sequential voice pair was different to the sentence spoken by the target voice (Experiment 7, Chapter 10).

10.2.2 Method

10.2.2.1 Design

The experiment employed a one-way between subject's ANOVA. The between-subjects factor was experimental manipulation (Experiment 5a: the use of a 1-second inter-stimulus interval between presentation of the target voice and the sequential voice pair, Experiment 6: the use of a 5-second inter-stimulus interval between presentation of the target voice and the sequential voice pair, and Experiment 7: using a different sentence in the sequential voice pair to the one previously spoken by the target voice). The dependant variable measured was mean percentage of errors made (i.e., percentage of time listeners chose the distractor voice instead of the target voice).

10.2.2.2 Participants

A total of 30 individuals (15 males; 15 females) contributed to Experiment 5a (Chapter 8). The ages of the participants ranged from 18 to 27 years old (M = 21.03 years, SD = 2.09 years). A total of 30 individuals (15 males; 15 females) contributed to Experiment 6 (Chapter 9). The ages of the participants ranged from 18 to 26 years old (M = 21.53 years, SD = 2.98 years). Finally, a total of 30 individuals (15 males; 15 females) contributed to Experiment 7 (Chapter 10). The ages of the participants ranged from 19 to 29 years old (M = 23.52 years, SD = 2.17 years).

10.2.2.3 Stimuli and Materials

The stimuli and materials were those used for the F0 condition in Experiment 5a (Chapter 8, Section 8.1.2.3), Experiment 6 (refer to Chapter 9, Section 9.1.2.3), and Experiment 7 (refer to Chapter 10, Section 10.1.2.3).

10.2.2.4 Procedure

The procedure was identical across all of the experimental manipulations, with the exception of there being only a 1-second inter-stimulus interval between presentation of the target voice and the sequential voice pair in Experiment 5a (refer to Chapter 8, Section 8.1.2.4), a 5-second inter-stimulus interval between presentation of the target voice and the sequential voice pair in Experiment 6 (refer to Chapter 9, Section 9.1.2.4), and a different sentence spoken in the sequential voice pair to the one previously spoken by the target voice in Experiment 7 (refer to Chapter 10, Section 10.1.2.4).

10.2.2.5 Analyses

The results were analysed using a one-way between-subjects ANOVA. A Bonferroni correction was applied to the simple main effects, which were conducted using independent samples *t*-tests.

10.2.3 Results

10.2.3.1 Fundamental Frequency (F0)

The mean matching errors scores were entered in a one-way ANOVA for the between-subjects factor of experimental manipulation (exploring the accentuation effect with a 1-second inter-stimulus interval between presentation of the target voice and the sequential voice pair (Experiment 5a), increasing the inter-stimulus interval to 5-seconds (Experiment 6), and using a different sentence in the sequential voice pair to the one previously heard (Experiment 7)). This revealed a significant main effect of experimental manipulation, $F(2, 87) = 4.59$, $p < .05$, $\eta_g^2 = 0.10$. Figure 10.5 shows that there was no difference in the amount of matching errors made overall between Experiment 5a when a 1-second inter-stimulus interval between presentation of the target voice and the sequential voice pair was used ($M = 15.80$, $SD = 7.39$)

and in Experiment 6 when the inter-stimulus interval between presentation of the target voice and the sequential voice pair was increased to 5-seconds (*M* = 17.19, *SD* = 11.31), *t*(58) = -.60, *p* > .05, *d* = 0.17[21]. In contrast, significantly more matching errors were made overall in Experiment 7 when the sentence used in the sequential voice pair was different to the sentence previously spoken by the target voice (*M* = 22.09, *SD* = 5.61) than in Experiment 5a when the sentence used in the sequential voice pair was the same as the sentence spoken by the target voice (*M* = 15.80, *SD* = 7.39), *t*(58) = -.60, *p* < .05, *d* = 0.78. More matching errors were also made overall in Experiment 7 when the sentence used in the sequential voice pair was different to the sentence spoken by the target voice (*M* = 22.09, *SD* = 5.61) than in Experiment 6 when the sentence used in the sequential voice pair was the same as the sentence spoken by the target voice, and when the inter-stimulus interval between presentation of the target voice and the sequential voice pair was increased. However, this just failed to reach significance, *t*(58) = -3.71, *p* = .08, *d* = 0.60.

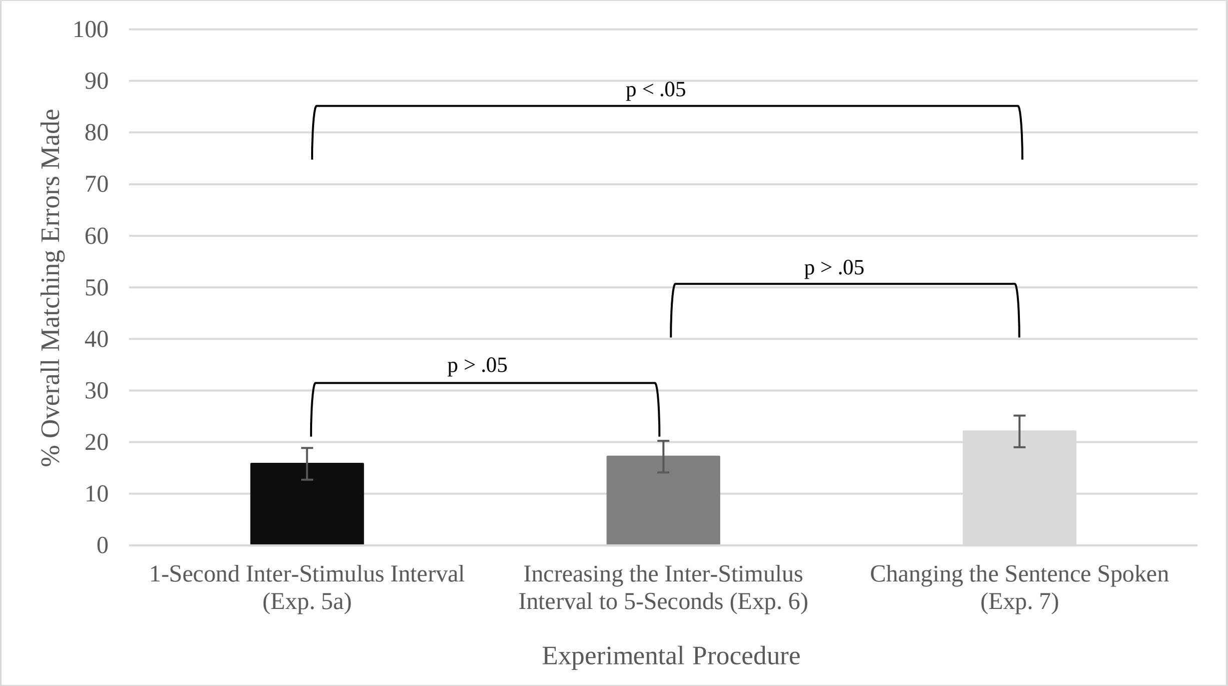


Figure 10.5: Mean percentage of matching errors made overall (i.e., chose distractor voice instead of target voice) for the three experimental manipulations (i.e., Experiment 5a, 6, and 7). 95% confidence intervals are also shown.

[21] Cohen’s *d* values were determined by calculating the mean difference between the two groups, and then dividing the result by the overall pooled standard deviation from all conditions (8.10).

10.2.4 Discussion

Experiment 8 investigated whether the amount of matching errors made overall between the three recognition memory experiments (Experiment 5a, Experiment 6, and Experiment 7) were different when manipulations in F0 were made. The findings indicated that there was no difference in matching errors made overall when a 1-second inter-stimulus interval was used (Experiment 5a, Chapter 8) and when a 5-second inter-stimulus interval was used (Experiment 6, Chapter 9). Therefore, the findings are in line with the proposed hypothesis. The findings also indicated that listeners made fewer matching errors overall when the content of the sentence in the sequential voice pair was the same as the sentence spoken by the target voice (Experiment 5a, Chapter 8) compared to when the content of the sentence in the sequential voice pair was different to the sentence spoken by the target voice (Experiment 7, Chapter 10). Therefore, the findings are in line with the proposed hypothesis. There was no difference in matching errors made overall in Experiment 7 (Chapter 10) and Experiment 6 (Chapter 9) when the sentence used in the sequential voice pair was the same as the sentence spoken by the target voice and the inter-stimulus interval between presentation of the target voice and the sequential voice pair was increased to 5-seconds, though this just failed to reach significance.

The finding that the amount of matching errors made overall was similar in Experiment 5a and Experiment 6 suggests that, at least for synthesised voices, recognition performance for voices may not be directly dependent on the time course of echoic memory. Recognition tasks are typically easier at shorter retention intervals (i.e., at a 1-second inter-stimulus interval compared to a 5-second inter-stimulus interval) because the acoustic trace is stronger. However, the findings presented here suggest that accurate recognition does not depend on being able to compare high quality auditory representations in memory, and that increasing the interval does not increase the load on auditory sensory memory. Rather, the findings suggest that the representation of vocal stimuli in auditory memory may be retained for periods longer than 3 to 5 seconds, and may leave a stronger trace in memory than non-vocal auditory stimuli. The present findings offer support to others that have found that the length of the inter-stimulus interval did not affect matching accuracy (Smith, Dunn, Baguley, and Stacey, 2016). Furthermore, it is possible that auditory acoustic information about F0 of the voice may be retained in memory for periods longer than five seconds, and for periods when the acoustic trace has supposedly degraded and is weak (Glanzer & Cunitz, 1966; Lu, Williamson & Kaufman, 1992; Treisman, 1964; Wickelgren, 1969).

The amount of recognition errors made overall in Experiment 7 was significantly higher to those made in Experiments 5a, offering support to others who have found recognition tasks to be typically harder when a different sentence is spoken to the one previously heard (e.g., Gliskey et al., 2001; Reid & Craik, 1995; Winograd, Kerr, & Spence, 1984). This suggests that, at least for synthesised voices, recognition performance may be dependent on the content of the spoken message and that listeners may be using the content of the sentence to help aid the recognition process when manipulations in F0 are made. In Experiment 7, listeners are forced to rely on acoustic properties of the voice (in this case, F0) rather than the content when a different sentence is used in the sequential voice pair to the sentence previously heard by the target voice. Conversely, in Experiment 5a, listeners are able to use similarities in elements of the spoken sentence *and* F0 to help aid recognition, and essentially making it easier for listeners to match voices in Experiment 5a than in Experiment 7.

10.2.4.1 Summary Conclusions

The results from the present experiment suggest that listeners make fewer matching errors overall when the same sentence is used in the sequential voice pair as the previously heard target voice compared to when a different sentence is used in the sequential voice pair to the previously heard target voice. Therefore, it is likely that listeners use both elements of the spoken message and F0 to aid the recognition process. However, there was no difference in errors made overall when a 1-second inter-stimulus interval was used and when a 5-second inter-stimulus interval was used between presentation of the target voice and the sequential voice pair. Thus, it appears that recognition performance for voices may not be directly dependent on the time course of echoic memory. Rather, the findings suggest that the representation of vocal stimuli in auditory memory may be retained for periods longer than 3 to 5 seconds, and may leave a stronger trace in memory than non-vocal auditory stimuli.

CHAPTER 11. SUMMARY OF FINDINGS AND GENERAL DISCUSSION

11. Introduction

This chapter summarises the main findings from the six experiments reported in this thesis. It also discusses the potential applied implications of these findings and suggests some directions for future research.

11.1 Summary of Aims

This thesis investigated the effect of manipulations in either F0 or speech rate on voice recognition performance and perceptions about the speaker's identity, sex, and age. Studies on the extent to which listeners can judge a speaker's physical characteristics are common in the voice literature (e.g., Hartman & Danhauer, 1976; Kreiman & Sidtis, 2013; Skuk & Schweinberger, 2013; Smith & Patterson, 2005), with research tending to suggest that manipulations in F0 are effective at changing the perceived identity (e.g., Kuwabara & Takagi, 1991; Lavner et al., 2000; Sell et al., 2015), sex (e.g., Assman et al., 2006; Coleman, 1971; Gelfer & Bennet, 2012; Gelfer & Mikos, 2005; Harnsberger et al., 2008; Hillenbrand & Clark, 2009; Whiteside, 1971), and age (e.g., Linville & Fisher, 1985; Shipp et al., 1992; Shrivastav et al., 2003; Smith & Patterson, 2005; Waller & Eriksson, 2016) of the speaker. For speech rate, research suggests that manipulations in speech rate are effective at changing the perceived age of the speaker (Harnsberger et al., 2008; Hartman & Danhauer, 1976; Ptack & Sander, 1966; Ryan & Burk, 1974; Shipp et al., 1992; Shrivastav et al., 2003; Waller & Eriksson, 2016). There is no clear consensus in the literature on whether manipulations in speech rate are likely to change perceptions of the identity or the sex of the speaker. The overall picture is still somewhat unclear, and previous research has presented contradictory findings (e.g., Brown, 1981; Byrd, 1992; 1994; Whiteside, 1996; Pepiot, 2014). There are also several methodological issues with the research that currently exist that limit the applicability of the findings. For example, some studies have used only one voice (e.g., Brown, 1981; Gaudrain et al., 2009), others have used familiar speakers rather than unfamiliar speakers (e.g., Lavner et al., 2000; Kuwabara & Takagi, 1991), and some have manipulated vowels or syllables rather than words or full sentences (e.g., Bennett & Montero-Diaz, 1982; Coleman, 1971; Schwartz & Rine, 1968; Lavner et al., 2000; Linville & Fisher, 1985; Whiteside, 1998).

Few researchers have investigated the role of manipulations in F0 or speech rate on recognition performance for the voice. The studies that do exist on this topic have demonstrated that manipulations in F0 or speech can lead to accentuation effects in voice memory (Mullenix et al., 2010; Stern et al., 2007). As a result, listeners appear to exaggerate the representation of a target voice in terms of F0 or speech rate, and mistakenly remember it as being higher or lower in F0, or faster or slower in speech rate, than the voice they originally heard (Mullenix et al., 2010; Stern et al., 2007). However, there are several methodological issues with the work that make it difficult to determine the relevance of the findings. Only one male voice was used making it difficult to determine whether the findings are generalisable to other voices and to female voices. Furthermore, the manipulations in F0 or speech rate fell outside the typical male and female F0 and speech rate ranges in everyday situations. This is important given that it is highly unusual to hear voices outside of the typical male and female range in everyday situations. There are also other factors that might lead to accentuation effects in voice memory, including whether voices are presented with and without a delay, and whether the same, or different sentences, are used. At present however, no research has explored these ideas further.

Over a series of experiments, this thesis has attempted to resolve these issues and unanswered questions, as well as advancing on the existing literature. The section that follows briefly summarises the main findings and conclusions of these experiments.

11.2 Summary of Findings and Main Conclusions

Experiments 2, 3, and 4 set out to determine the extent to which manipulations in F0 or speech rate affect perceptions about the identity, sex, and age of the speaker across a series of 2AFC perceptual voice discrimination tasks. Experiment's 5, 6, and 7 set out to determine whether manipulations in F0 or speech rate affect recognition performance for the voice, and if so, whether the pattern of findings identified were attributable to the accentuation effect. These experiments also explored whether any biasing affecting performance was dependent on whether the voices were presented with and without a delay, and whether the same or a different sentence was presented to the one that was previously heard.

Experiment 2 (Chapter 5) investigated whether manipulations in F0 or speech rate affect perceptions of speaker identity. The listeners were presented with an original voice and a manipulated version of the voice in either F0 or speech rate, and asked to decide whether the voices were the same identity or a different identity. The results suggested that manipulations in F0 or speech rate increased uncertainty about the identity of the speaker, though listeners were more robust to changes in speech rate than they were to changes in F0. It was concluded that F0 is more directly related to speaker identity than speech rate, and that listeners rely on F0 more than they do on speech rate when making decisions about the identity of the speaker.

Experiment 3 (Chapter 6) investigated whether manipulations in F0 or speech rate affect perceptual of speaker sex. The listeners were presented with one of the six original voices and the manipulated versions of the voices and asked to decide whether the voice they heard was male or female. The results suggested that manipulations in F0 increased uncertainty about speaker sex, but shifts in speech rate did not. Male voices were accurately perceived as male irrespective of the direction of manipulation in F0. However, for female voices, decreasing F0 increased the uncertainty of speaker sex (i.e., the voices were more likely to be perceived as male rather than female). It was concluded that F0 is more directly related to speaker sex than speech rate, and that listeners rely on F0 more than they do on speech rate when making decisions about the sex of the speaker.

Experiment 4 (Chapter 7) investigated whether manipulations in F0 or speech rate affect perceptions about speaker age. The listeners were presented with one of the six original voices and the manipulated versions of the voices and were asked to freely estimate the age of the speaker. The results suggested that increasing either F0 or speech rate resulted in both male and female voices as sounding younger, whereas decreasing either F0 or speech rate lead to listeners perceiving the voices as sounding older. However, some discrepancy appeared to exist between listeners expectations about speakers of different ages and the vocal characteristics that actually exist. For male voices, decreasing F0 lead to the voices as being perceived as sounding older regardless of any changes in F0 that actually occur. Furthermore, for both male and female voices, listeners perceive changes in speech rate occurring at a much younger age than they actually do. It was concluded that both F0 and speech rate are important cues for estimating speaker age.

Experiment 5a and 5b (Chapter 8) investigated whether manipulations in F0 or speech rate affect recognition performance for the voice, and whether the findings are attributable to the accentuation effect. Using a 2AFC procedure, the listeners were asked to recognise a previously heard target voice from a sequential voice pair that included the target voice and a manipulated version of the voice. A 1-second inter-stimulus interval was used between presentation of the target voice and the sequential voice pair. The results showed that manipulations in F0 or speech rate did increase matching errors for the target voice, however, there was no evidence for an accentuation effect. Specifically, for voices manipulated in F0, there was an increase in the selection of voices higher in F0 compared to voices lower in F0 for high, moderate, and low F0 target voices. For voices manipulated in speech rate, there was an increase in the selection of voices faster in speech rate compared to voices slower in speech rate, but only for slow speech rate target voices.

Experiment 6 (Chapter 9) investigated whether increasing the inter-stimulus interval lead to accentuation effects in voice memory when voices were manipulated in F0. The procedure was identical to that in Experiment 5a, with the exception of a 5-second inter-stimulus interval between presentation of the target voice and the sequential voice pair. This experiment was designed to push the target voice out of the range, or at least to the very limits, of sensory memory. Overall the pattern of results observed in Experiment 6 were similar to those observed in Experiment 5a. It was concluded that listeners are no more likely to rely on self-generated categorical information about the voice at the time of encoding to aid recognition with an increased inter-stimulus interval.

Experiment 7 (Chapter 10) investigated whether a different sentence used in the sequential voice pair to the one previously spoken by the target voice lead to accentuation effects in voice memory when voices were manipulated in F0. The procedure was identical to that in Experiment 5a, with the exception of a different sentence being spoken as part of the sequential voice pair. Overall, the pattern of results observed in Experiment 7 were similar to those observed in Experiment 5a and 6. It was concluded that listeners were no more likely to rely on self-generated categorical information about the voice at the time of encoding to aid recognition when a different sentence is spoken to the one previously spoken by the target voice.

Experiment 8 (Chapter 10) set out to determine whether the number of matching errors made overall were different for the three recognition memory experiments (Experiment 5a, Experiment 6, and Experiment 7) when manipulations in F0 were made. This was to determine whether accurate matching decisions rely on high quality representations temporarily stored in echoic memory (Experiment 5a, Chapter 8), and whether the ability to make accurate decisions diminishes as the inter-stimulus interval increases (Experiment 6, Chapter 9). The experiment also explored whether listeners were more accurate at matching voices when the sentences spoken in the sequential voice pair were the same as the sentence spoken by the target voice (Experiment 5a, Chapter 8), compared to when the sentences spoken in the sequential voice pair were different to the sentence spoken by the target voice (Experiment 7, Chapter 10). The findings suggest that there was no difference in matching errors made for voices when the inter-stimulus interval between presentation of the target voice and the sequential voice pair was increased to 5-seconds (Experiment 6, Chapter 9) compared to only 1-second. Thus, the representation of vocal stimuli in auditory memory may leave a stronger trace in memory than non-vocal auditory stimuli. The findings also suggested that listeners made fewer matching errors overall when the target and voice pairs contained the same sentence than when they were different. Thus, listeners are likely to use patterns identified in F0 of the sentence spoken to help aid the recognition process.

11.2.1 Main Conclusions

In consideration of the findings discussed above, the following main conclusions can be drawn, each of which offer an independent contribution to knowledge:

- **(1)** Fundamental frequency (F0) provides important information about the identity, sex, and age of unfamiliar voices. Manipulations in F0 are likely to affect perceptions of speaker identity, sex, and age. Speech rate provides important information about the age of the speaker. Manipulations in speech rate are likely to affect perceptions of speaker age.

- **(2)** Manipulations in F0 or speech rate affect recognition performance for unfamiliar voices. However, the pattern of errors identified are difficult to explain using the accentuation effect.

- **(3)** Accentuation errors are no more likely to occur for unfamiliar voices when they are manipulated in F0 and the inter-stimulus interval between presentation of a target voice and a sequential voice pair is increased. Accentuation errors are also no more likely to occur for unfamiliar voices when they are manipulated in F0 and a different sentence is spoken in the sequential voice pair to the one previously spoken by the target voice.

- **(4)** Recognition errors made for unfamiliar voices are similar when a 1-second inter-stimulus interval is used and when a 5-second inter-stimulus interval is used between presentation of a target voice and a sequential voice pair. Thus, the representation of vocal stimuli in auditory memory may leave a stronger trace in memory than non-vocal auditory stimuli.

- **(5)** Listeners make fewer matching errors for unfamiliar voices when the target and voice pairs contain the same sentence than when they are different. Thus, listeners are likely to use patterns identified in F0 of the sentence spoken to help aid the recognition process.

11.3 Research Questions

Three overarching research questions were outlined in Chapter 3. Each of these will be addressed in turn, drawing together the evidence from the experiments carried out to help facilitate a detailed consideration of the results. It will expand on the main conclusions referred to above by explaining how the findings contribute further to our existing knowledge.

11.3.1 Research Question 1: Do manipulations in fundamental frequency (F0) and speech rate affect perceptual judgments about the paralinguistic characteristics of the speaker, and if so, how do they affect them? Specifically, how do manipulations in F0 or speech rate affect perceptions of,

a) speaker identity?

b) speaker sex?

c) speaker age?

Studies of the extent to which listeners can judge a speaker’s characteristics using the voice alone are fairly common in the literature (Kreiman & Sidtis, 2013). However, despite this, the overall picture is still somewhat unclear with research presenting contradictory findings as to whether F0 or speech rate change perceptions about characteristics of the speaker. Furthermore, there are several methodological issues with the work so far that make it difficult to determine the relevance of the findings. Experiments 2 (Chapter 5), 3 (Chapter 6), and 4 (Chapter 7) addressed this gap in the literature by testing whether manipulations in F0 or speech rate affect perceptions of the identity, sex, and age of the speaker.

11.3.1.1 Perceptions of Speaker Identity

Few researchers have attempted to determine the contribution of F0 or speech rate in perceptions of speaker identity. What work that does exist tends to suggest that for F0, manipulations in F0 are important when making perceptual judgements about the identity of the speaker (e.g., Kuwabara & Takagi, 1991; Lavner, Gath, & Rosenhouse, 2000; Sell et al., 2015). Research tends to suggest that greater manipulations in F0 increase the likelihood that the voice is perceived as being a different speaker (i.e., a different identity) (e.g., Kuwabara & Takagi, 1991; Lavner, Gath, & Rosenhouse, 2000). However, others have found that whilst changes in F0 do affect the ability of the listener to correctly identify speakers, listeners are still able to consistently perform at a high level, and always above chance (e.g., Sell et al, 2015). Moreover, susceptibility to manipulations in F0 have been found to be dependent on the specific speaker (e.g., Lavner, Gath, & Rosenhouse, 2000; Sell et al., 2015), suggesting that individual speaker variation is important when determining the role of F0 in speaker identity. Nevertheless, several methodological issues, including the type of procedure employed (i.e., judgements of speaker similarity rather than speaker identity were made in Gaudrain et al., 2009) and the stimuli set used (i.e., familiar voices rather than unfamiliar voices in Kuwabara & Takagi, 199; spoken vowels rather than words or sentences in Lavner et al., 2000; and male voices rather than female voices in Sell et al., 2015), make it difficult to determine the relevance of these findings.

Experiment 2 (Chapter 5) addressed these methodological issues by using a 2AFC same/different identity task, and a larger set of voices which were unfamiliar to the listener and included both male and female speakers. The findings in Experiment 2 (Chapter 5) were overall consistent with the research carried out so far, and in line with the original predictions made. Manipulations in F0 did affect perceptions about the identity of the speaker. Specifically, larger manipulations in F0 (i.e., 10%) increased the likelihood that the voices would be perceived as a different identity than the original version of that voice. Importantly, the findings suggested that performance fell below chance when larger manipulations in F0 were made. As previously noted, some studies have found listeners to consistently perform above chance (e.g., Sell et al., 2015). The findings from Experiment 2 (Chapter 5) did not support this view. One possible explanation for the difference in the findings is that Sell et al. (2015) made only very small manipulations in F0. This made it difficult to determine whether manipulations that were larger in F0, and were also achievable within the human vocal range, would change perceptions of speaker identity. The manipulations made in F0 for the voices in this thesis all remained within the typical male and female F0 ranges of voiced speech. Therefore, the findings of Experiment 2 (Chapter 5) suggest that manipulations in F0 that are achievable for both male and female speakers can change the perceived identity of the speaker for unfamiliar voices.

The findings from Experiment 2 (Chapter 5) did not support previous work that has shown susceptibility to manipulations in F0 to be dependent on the specific speaker (e.g., Lavner, Gath, & Rosenhouse, 2000; Sell et al., 2015). Rather, manipulations in F0 had a similar effect for all voices, suggesting that F0 is a cue that is likely to be utilised by listeners in determining the identity of all speaker's (i.e., the cues used are not dependent on the specific speaker). Furthermore, a similar pattern of findings was identified for both male and female voices, suggesting that the effect of manipulations in F0 on the perceived identity of the speaker found for male voices in the literature (e.g., Lavner et al., 2000; Sell et al., 2015) is found for female voices too. It is also important to emphasise that the stimuli used in Experiment 2 (Chapter 5) were more similar to those voices that are heard in the real-world, using real sentences instead of vowel sounds and nonsense words. Experiment 2 (Chapter 5) therefore provides a more robust and conclusive set of findings about the effect of manipulations in F0 on perceptions of speaker identity than what is currently available.

To the best of this authors knowledge, there is no work in the literature exploring the effect of manipulations in speech rate on perceptions of speaker identity, and only one study has explored the effect of manipulations in speech rate on perceptions of speaker similarity (e.g., Brown, 1981). Brown (1981) found that manipulations in speech rate did affect similarity judgements. Specifically, greater manipulations in speech rate lead to listeners perceiving the voice as sounding less similar to a control voice (i.e., the original version of that voice). However, this work cannot establish whether manipulations in speech rate also affect perceptions of speaker identity. Furthermore, Brown (1981) used only one male voice making it difficult to determine whether the results are generalisable to female voices, or indeed to any other voice. The findings of Experiment 2 (Chapter 5) suggest that manipulations in speech rate do increase the uncertainty of the identity of the speaker, however listeners are largely robust to these changes and are always able to correctly identify (above chance) the speaker. A similar pattern of findings was observed for all voices, suggesting that the findings are generalisable to more than one voice and to both male and female voices.

Given the above, it is likely that F0 is more directly related to speaker identity than speech rate, and that listeners rely on F0 more than they do on speech rate when making decisions about the identity of the speaker. The discrepancy in the findings for F0 and speech rate might be because within-speaker variations in speech rate are more variable compared to F0, which is relatively stable in everyday situations (e.g., Mullenix et al., 2010; Stern et al., 2007). Thus, it is likely that listeners are more familiar with speech rate variability and hence, are more robust to the changes that occur as a result of the manipulations made. Alternatively, we might ignore speech rate in favour of other cues (namely F0) when assessing identity. Differences between F0 and speech rate might also exist because of the type of information that is typically portrayed in these cues. F0 is strongly determined by the physiological and anatomical structures of the vocal tract (Fant, 1966). Consequently, it is likely that F0 is more directly related to speaker identity than speech rate. In contrast, speech rate is more useful for conveying emotional state, motivations, and the intention of the speaker (Livingstone, 2015; Sbattella, Colombo, Rinaldi, Tedesco, Matteucci & Trivilini, 2014). Therefore, the findings in Experiment 2 (Chapter 5) have established that there are important differences between the acoustic cues of the voice, and that their likely effect on perceptions of speaker identity will be different.

11.3.1.2 Perceptions of Speaker Sex

The existing literature on whether manipulations in F0 affect perceptions of speaker sex are more prevalent than they are for speaker identity. However, the overall pattern is still somewhat unclear. The research tends to suggest that manipulations in F0 are likely to change the perceived sex of the speaker (e.g., Assman et al., 2006; Coleman, 1971; Gelfer & Mikos, 2005; Hillenbrand & Clark, 2009; Whiteside, 1998). In the main, the research suggests that there does appear to be some evidence to suggest a male-advantage in the perception of speaker sex, where listeners are more accurate at determining the sex of the speaker for male voices than they are for female voices (e.g., Coleman, 1971; Gelfer & Mikos, 2005; Owren, Berkowitz, and Bachorowski, 2007). Specifically, the presence of critical features of maleness (i.e., low F0, low formant values) virtually guarantees that the speaker is an adult male. However, their absence does not unequivocally imply that the talker is an adult female (Owren, Berkowitz, & Bachorowski, 2007). Thus, manipulations in F0 have been found to affect perceptions of speaker sex for female speakers more than they do for male speakers (e.g., Coleman, 1971; Gelfer & Mikos, 2005; Owren, Berkowitz, and Bachorowski, 2007). In contrast, others have found that female speakers are accurately perceived as female even with an F0 in the gender ambiguous range, or in the typical male range (e.g., Hillenbrand & Clark, 2009; Pausewang, Gelfer, & Bennett, 2012). Furthermore, for male speakers, listeners are less accurate at perceiving the sex of the speaker with an F0 in the typical female range (Gelfer & Bennett, 2013). However, the evidence is still somewhat limited and there are several methodological issues with the research carried out so far. Indeed, studies have typically only used isolated vowels (Coleman, 1971; Gelfer & Mikos, 2005; Kreiman & Sidtis, 2013;

Whiteside, 1998), making it difficult to determine whether similar findings could still be identified for spoken sentences.

Experiment 3 (Chapter 6) addressed these methodological issues by using a larger set of participants, and a set of stimuli where complete sentences were spoken. The findings are therefore likely to be more robust and informative about the effect of manipulations in F0 on perceptions of speaker sex than what is currently available. The findings in Experiment 3 (Chapter 6) were consistent with much of the existing literature (e.g., Assman et al., 2006; Coleman, 1971; Gelfer & Mikos, 2005; Hillenbrand & Clark, 2009; Whiteside, 1998). Manipulations in F0 did affect perceptions of speaker sex. However, this was only apparent for the female voices. Specifically, those female voices that were decreased in F0 and fell in the typical male F0 range, were more likely to be perceived as male than they were female. In contrast, male voices were accurately perceived as being male, even when male voices were increased in F0. The findings in Experiment 3 (Chapter 6) therefore offer further support to previous work that has identified a male advantage in the perception of speaker sex, where listeners are more accurate at determining the sex of the speaker for male voices than they are for female voices (e.g., Assman et al., 2006; Gelfer & Mikos, 2005). Furthermore, the findings in Experiment 3 (Chapter 6) supported the view that there appears to be a bias towards selecting the speaker as being male when any critical feature of maleness (i.e., low F0, low formant values) is present (Coleman, 1976). Indeed, those female voices in Experiment 3 (Chapter 6) that still fell in the typical female F0 range, were also perceived as male for some percentage of the time. Observations of vowel formant frequencies for these voices showed that they fell in the typical male formant frequency range for voiced speech (refer to Chapter 4, Section 4.1.1.1). Despite this however, female voices with formant frequency values that fell in the typical male range were still more likely to be perceived as female rather than male. In contrast, those female voices with an F0 that fell in the typical male F0 range were more likely to be perceived as male rather than female. These findings lend support to the suggestion that there is a bias towards selecting the speaker as being male when any critical feature of maleness is present, and also suggests that F0 is likely to be the more robust cue in determining speaker sex (Coleman 1971).

It should be noted that the male voices that were increased in F0 did not fall in either the gender ambiguous range or the typical female F0 range for voiced speech, making it difficult to determine whether perceptions of speaker sex for male voices would change if they did. Nevertheless, increasing F0 did not change the perceived sex of the speaker for male voices at all. Moreover, male voices that fell close to the gender ambiguous range were no more likely to be perceived as being female than those that fell in the lower male F0 range. In contrast, female voices that fell closer to the gender ambiguous range increased the uncertainty of the sex of the speaker. The findings in Experiment 3 (Chapter 6) suggest that it is unlikely that listeners would have perceived male voices that were increased in F0 and fell in the female F0 range as female, or at the very least for a greater percentage of the time than they would be perceived as male. Further research would be required however to determine whether this was indeed the case.

Potential male-female differences in speech rate have also been investigated in the literature, although somewhat less extensively than they have for F0. Research has shown that people typically believe females speak at a faster rate than males (Weirich & Simpson, 2014). This belief is quite pervasive and has been given credence in the scientific literature, with some even making the unsubstantiated claim that females speak on average faster than males (Brizendine, 2006). In fact, and contrary to pervasive popular opinion, research suggests that males have a faster speaking rate than females (e.g., Byrd, 1992, 1994; Pepiot, 2014; Whiteside, 1996), though to this authors knowledge, no one has yet explored speech rate as a cue to speaker sex. The results of Experiment 3 (Chapter 6) suggest that manipulations in speech rate are unlikely to change perceptions of speaker sex.

The discrepancy between the findings for F0 and speech rate are likely due to the information obtained by the listener using F0 or speech rate. Humans exhibit large sexual dimorphism in vocalisations and vocal anatomy (Puts, Apicella & Cardenas, 2011) that occur because of hormonal changes during puberty (e.g., Fitch & Giedd, 1999; Harries et al., 1998; Newman et a., 2000; Puts et al., 2007). Because of this, males have large vocal folds and vocal tracts than females, giving rise to markedly lower F0's in male voices (Baken, 1987; Titze, 1994). This is likely to make F0 a particularly decisive cue in the identification of speaker sex (Pepiot, 2015; Perry, Ohde, & Ashmeand, 2001). In contrast, speech rate is useful for conveying emotional state, motivations, and the intention of the speaker (Livingstone, 2015; Sbattella, Colombo, Rinaldi, Tedesco, Matteucci & Trivilini, 2014). Furthermore, any differences that have been identified in the speech rate of male and female speakers can be small (Yuan et al., 2006), or have not been found at all (Block & Killen, 1996; Robb et al., 2004. Consequently, speech rate is unlikely to offer a decisive cue to speaker sex.

11.3.1.3 Perceptions of Speaker Age

Research typically suggests that speech rate, and to a lesser extent F0, are important in estimating perceptions of speaker age (e.g., Smith & Patterson, 2005; Smith, Walters, & Patterson, 2007; Waller & Eriksson, 2016). The prevalent finding in the existing literature is that increasing F0 and speech rate leads to male and female voices being perceived as sounding younger, whereas decreasing F0 and speech rate leads to voices being perceived as sounding older (e.g., Braun & Rietveld, 1995; Hartman & Danhauer, 1976; Horii & Ryan, 1981; Linville & Fisher, 1985; Ptack & Sander, 1966; Shipp et al., 1992). However, changes in speech rate are often perceived as occurring at a much younger age than they actually do, suggesting that discrepancies might exist between listener's expectations about speakers of different ages and the vocal characteristics that actually exist (Hartman & Danhauer, 1976). Listeners are also found to consistently associate lower F0 with old age in both male and female speakers despite reported increases in F0 with age in males, suggesting the presence of vocal stereotyping by listeners regarding F0 and speech rate with age (Hartman & Danhauer, 1976). However, the evidence is still somewhat limited and there are several methodological issues with the research carried out so far. Indeed, research has typically used only male voices (e.g., Braun & Rietveld, 1995; Horri & Ryan, 1981; Shipp et al., 1992; Ryan & Burk, 1974), making it difficult to determine whether the same cues are used to make age estimates for female speakers. Studies have also used vowel sounds (Linville & Fisher, 1985), making it difficult to determine whether similar findings can be identified for spoken sentences. Furthermore, studies have asked listeners to attribute vocal characteristics to different groups of speakers (Ptack & Sander, 1966) rather than manipulating voices in either F0 or speech rate and asking people to determine their age. It is often acknowledged that experimental work in which the parameter of interest is manipulated constitutes much harder casual evidence for the effect of acoustic cues on age estimations (Waller, Eriksson, & Sorqvist, 2015). This is because any changes in the age estimates can be compared against a control voice (e.g., an original voice) to determine the extent to which manipulations in the cue affect speaker age.

Experiment 4 (Chapter 7) addressed these methodological issues by using both male and female voices, speaking complete sentences. Experiment 4 (Chapter 7) also manipulated F0 or speech and asked listeners to estimate the age of the speaker. The findings showed that manipulations in both F0 and speech rate affected perceptions of speaker age. Specifically, increasing F0 and speech rate lead to the listeners perceiving the voices as sounding younger than the original voices, whereas decreasing F0 and speech rate lead to the listeners perceiving the voices as sounding older than the original voices. This is consistent with much of the existing literature that has found listeners to associate a lower F0 and a slower speech rate with older age (e.g., Hartman & Danhauer, 1976; Waller & Eriksson, 2016). The results also lend further support to the suggestion that some stereotyping of the vocal characteristics for male voices might exist. Indeed, for female voices, the findings follow a similar pattern to the differences observed in female voices where F0 continues to drop from childhood to adulthood, and through to older age (refer to Chapter 3, Section 3.1.3.1.1). For male voices, whilst the findings are also in line with previous research (e.g., Hollien, et al., 2003; Hollien et al., 2008; Smith et al., 2007; Smith & Patterson, 2005; Winkler, 2007), they do not follow the pattern observed for male voices, where F0 decreases from childhood through to adulthood and into middle age, but then rises again into older age (refer to Chapter 3, Section 3.1.3.1.1).

Experiment 4 (Chapter 7) also found that male voices were perceived as sounding significantly older than female voices. This is consistent with the notion that, regardless of whether a voice is male or female, voices lower in F0 are typically perceived as sounding older than voices higher in F0, and is consistent with previous work that has found listeners to associate a lower F0 with older age (e.g., Hartman & Danhauer, 1976; Waller & Eriksson, 2016). Consequently, because male voices have a lower overall mean F0 compared to female voices, they are perceived as sounding older than female voices.

For speech rate, the findings were somewhat in line with the differences observed in male and female voices, and are consistent with previous literature suggesting that listeners are likely to be perceived as sounding older when speech rate is decreased (e.g., Braun & Rietveld, 1995; Ptack & Sander, 1966; Ryan & Burk, 1974). However, actual changes reported suggest that speech rate increases and reaches its peak rate during the mid-40’s, before it gets progressively slower into older age (e.g., Jacewicz & Fox, 2010; Kowal et al., 1975; Walker et al., 1992) The findings in Experiment 4 (Chapter 7) suggest that even voices perceived as sounding younger than the mid 40’s were rated as sounding progressively older as speech rate was decreased. Thus, it is likely that listeners perceive changes in rate occurring at a much younger age than they actually do. This finding lends support to the suggestion that discrepancies might exist between listeners expectations about speakers of different ages and the vocal characteristics that exist, and is consistent with previous work that has identified a similar pattern of findings (Hartman & Danhauer, 1976).

Male voices were also perceived as sounding significantly older than female voices when they were manipulated in speech rate. This finding provides evidence that both F0 and speech rate cues are used to estimate speaker age. Indeed, it is likely that listeners are still using F0 to make estimations about speaker age even though voices were manipulated in speech rate rather than F0. Since male voices have a lower overall mean F0 compared to female voices, male voices are likely to still sound older than female voices when manipulations in speech rate are made. The evidence provided in Experiment 3 (Chapter 6) suggests that F0 is highly indicative of speaker sex, thus it is unlikely that listeners will ignore cues in F0 when male and female voices are heard.

It is important to emphasise that the stimuli set used in Experiment 4 (Chapter 7) were more realistic than the voices used in previous research (e.g., Coleman, 1971; Gaudrain et al., 2009; Kuwabara & Takagi, 1991; Lavner et al., 2000). This is because they are more similar to those voices that are heard in the real-world, using real sentences instead of vowel sounds. Furthermore, the procedure used provides much harder causal evidence for the acoustic cues on age estimations (Waller, Eriksson, & Sorqvist, 2015). Thus, the findings are likely to be more robust and informative about the effects of manipulations in F0 or speech rate on perceptions of speaker age than what is currently available in the existing literature.

11.3.1.4 Summary Conclusions

In summary, Experiment 2 (Chapter 5), 3 (Chapter 6), and 4 (Chapter 7) have identified that greater manipulations in F0 increase the likelihood of changing perceptions of the identity and age of the speaker. For female voices, decreasing F0 also increased the likelihood that the voices would be perceived as male. Manipulations in speech rate had little influence on the perceived identity or sex of the speaker. However, manipulations in speech rate did affect perceptions of speaker age. The findings have contributed further to knowledge by identifying those cues that are likely to change perceptions of the paralinguistic properties of the speaker. The experiments have addressed several methodological issues and have therefore provided a more robust and informative set of findings than those previously identified in the existing literature.

11.3.2 Research Question 2: Do manipulations in fundamental frequency (F0) and speech rate affect recognition performance for the voice, and if so, can the findings be explained using the accentuation effect?

Few researchers have considered the impact of manipulating acoustic cues of the voice on recognition performance (i.e., Mullenix et al. 2010; Stern et al., 2007). This is important because intra-speaker variation in a speaker's voice, whether unintentional or deliberate, can greatly reduce recognition performance (Reid & Duke, 1979). The studies that do exist on this topic found evidence for accentuation effects in voice memory for F0 where listeners mistakenly selected voices lower in F0 than the low F0 target voice, and voices higher in F0 than the high F0 target voice (Mullenix et al. 2010; Stern et al., 2007). However, there was no difference in the selection of higher or lower F0 distractor voices for moderate F0 target voices. For speech rate, listeners mistakenly selected voices slower in rate than the slow rate target voices only (Mullenix et al. 2010; Stern et al., 2007). Nevertheless, there was no difference in the selection of faster and slower rate distractor voices for moderate or fast rate target voices. However, given the few studies that exist on accentuation effects in relation to voices, the evidence is limited, and there are several methodological issues with the research carried out so far. Studies have typically only used one male voice, making it difficult to determine whether the results are generalisable to all voices, and whether the same acoustic cues are used to recognise female speakers. Manipulations in F0 or speech rate also fell considerably outside the F0 and speech rate ranges that are typical in the English-speaking population. It is important to keep manipulations within the typical ranges so that findings are generalisable to those voices that are heard in the real-world. Indeed, it is highly unusual to hear voices outside of the typical F0 and speech rate ranges in everyday situations.

Experiment 5a and 5b (Chapter 8) addressed these methodological issues by using a larger set of synthesised male and female voices. The target and distractor voices were also kept within the F0 and speech rate ranges that are typical in the English-speaking population. Sex of voice and listener sex were also included as measures in the experimental design. This was deemed appropriate because research has emphasised sex differences in verbal memory tasks, with females outperforming males (Herlitz, Nilsson, & Backman, 1997; Lewin, Wolgers, & Herlitz, 2001; McGivern et al., 1997). Others have also reported an own-gender bias for unfamiliar voices (Roebuck & Wilding, 1993). Experiment 5a and 5b (Chapter 8) suggested that manipulations in F0 or speech rate did affect recognition performance for the voice. However, the findings were inconsistent with previous work (Mullenix et al., 2010; Stern et al., 2007) and difficult to explain using the accentuation effect. Experiment 5a (Chapter 8) showed that for F0, there was an increase in the selection of voices higher in F0 compared to voices lower in F0 for high, moderate, and low F0 target voices. Experiment 5b (Chapter 8) showed that for speech rate, there was an increase in the selection of voices faster in speech rate compared to voices slower in speech rate for slow speech rate target voices. The findings of Experiment 5a and 5b (Chapter 8) suggest that listeners are susceptible to distortions in memory for F0 more so than they are for speech rate.

There are several possible explanations as to why listeners make more recognition errors when voices are paired with distractor voices higher in F0 compared to when they are paired with distractor voices lower in F0. First, it is possible that the listeners had difficulty discriminating between the frequencies of some of the voice pairs in the experiment. This is consistent with other research that has found it more difficult to discriminate between voices of higher frequencies compared to voices at lower frequencies (Moore, 1995). In Experiment 5a (Chapter 8), listeners may have made fewer errors identifying target voices when paired with distractor voices lower in F0 because they were more efficient at detecting the changes in frequency than when distractor voices were higher in frequency. This interpretation would account for why there was no effect of listener sex on errors made identifying target voices, because there is no reason to believe that the perceptual capabilities of the listener would differ substantially between male and female listeners. It would also explain why there was no difference in errors made for male and female target voices. Although female voices are higher in F0 than male voices, the findings are based upon a listener's ability to detect any *differences* in the frequencies of the voices in the voice pair, and this is independent of the frequency of the target voice itself.

It is also possible that listeners made more errors identifying target voices when paired with distractor voices higher in F0 compared to when they were paired with distractor voices lower in F0 because they a more like those voice heard in the real world. Indeed, research suggests that people are more likely to inflect the frequency of their voice upwards rather than downwards (in other words, they are more likely to increase, rather than decrease, the frequency of their voice when they speak) (e.g., Barbaranne, 1981; Fairbanks & Pronovost, 1939). For example, speakers typically use upward inflections when asking a question (Ching, 1982). Thus, the listeners in Experiment 5a (Chapter 8) might have been selecting distractor voices higher in F0 more often than distractor voices lower in F0 because they are more familiar with these types of utterances and it sounds like a more plausible version of the target voice (i.e., an inflected version of the target voice).

The finding that listeners were more likely to select distractor voices higher in F0 compared to distractor voices lower in F0 was particularly prevalent for the low F0 target voice condition in Experiment 5a (Chapter 8). This bias may have arisen because voices higher in F0 are perceived as less threatening than voices lower in F0. Research has shown that both male and female voices lowered in F0 are perceived as more dominant, threatening, and aggressive than the same voices raised in F0 (Bolinger, 1964; Borkowska & Pawlowski, 2011; Fraccaro, O’Connor, Re, Jones, DeBruine, & Feinberg, 2012; Jones, Feinberg, DeBruine, Little, & Vukovic, 2010; Morton, 1994; Ohala, 1984; Puts, Gaulin, & Verdonili, 2006). Furthermore, evidence tends to suggest that people will often exhibit avoidance type behaviour when exposed to aversive stimuli (Corr, 2013). Assuming that in Experiment 5a (Chapter 8), the voices lower in F0 would be rated as sounding more dominant and threatening than the voices higher in F0, listeners may have selected the higher voice of the pair because it sounded less dominant and less threatening. This would explain why an increase in the selection of higher F0 distractors was particularly prevalent for the low F0 target voice condition; because the voices were decreased in F0 sufficiently for the higher F0 voices in the pair to be perceived as less threatening to the listener. It would also account for why there was no effect of either sex of voice or listener sex; perceptions of dominance have been found to be equivalent for both male and female voices and male and female listeners (Jones, Feinberg, DeBruine, Little, & Vukovic, 2010).

Another possibility for why the selection of distractor voices higher in F0 compared to distractor voices lower in F0 was particularly prevalent for the low F0 target voice condition in Experiment 5a (Chapter 8), is that English voices lower in F0 for both males and females tend to co-occur with covariations in voice quality (e.g., Aberton, Howard, & Fourcin, 1989). A bias towards selecting the higher F0 distractor voices could reflect the unnaturalness of the voices lowered in F0 without a concomitant change in voice quality. Whilst the voices used in this experiment were rated as sounding natural, this issue might still remain even if naturally sounding voices were modified to have a lower F0.

Finally, it is worth noting that the naturalness ratings for the voices with higher F0 manipulations tended to yield slightly higher naturalness rating scores than those with lower manipulations (refer to Chapter 4, Section 4.3.4). One possible interpretation of this is that the listeners preferred the more natural sounding voices (i.e., the higher F0 manipulations) and were thus, more likely to select them. Unfortunately, because the naturalness ratings came from a different population to those in the 2AFC tasks reported here, it was not appropriate to formally test this possibility after the fact. However, given that the voices were generally perceived to be natural sounding across the board, and any differences observed between the voices were relatively small, are unlikely to have impacted upon the matching tasks. Thus, whilst it is a possibility that naturalness may have an effect, the question is unable to be resolved here.

There are several possible explanations as to why listeners selected voices faster in speech rate when slow speech rate target voices were presented. It is possible that the findings in Experiment 5b (Chapter 8) could be accounted for by the listener's level of familiarity of the voice heard. In natural speech, a person speaking more slowly is likely to be more hesitant, making more silent pauses or filled pauses (e.g., *um, er*). There are no silent or filled pauses in the synthesised voices used here. In Experiment 5b (Chapter 8), decreasing speech rate did affect the rate of continuous production but did not lead to increased pauses of any kind. It is therefore unlikely that the speech samples used were an entirely natural rendition of slower speech, at least of a type that listeners most typically hear. It is possible that at the lower margins of the speech rate manipulated samples (i.e., the slowest samples), but not elsewhere, the participants may have selected a faster voice in the pair because it sounded relatively more realistic.

Faster speaking voices might also sound more favourable when compared with slower speaking voices in the slow speech rate pairings. Indeed, research suggests that speech rates can influence a listener's perceptions of a speaker's personality and social skills. For example, faster speaking styles have been shown to be rated more favourably (Stewart & Ryan, 1982), and viewed as more competent and socially attractive than voices spoken at a slower rate (Street, Brady, & Putman, 1983). Slower speaking styles have also been identified as sounding weaker, less truthful, and less empathetic than voices spoken at a faster rate (Apple, Streeter, & Krauss, 1979). It is possible that listeners were more likely to select a faster voice in the pair because they preferred the sound of the voice. However, such selections may have been made only for the slow speech rate condition because these voices were slowed sufficiently for the faster rate voices in the pair to be rated more favourably, and thus selected by the listener. The above explanations would also account for why there was no effect of either sex of voice or listener sex on errors made identifying a target voice, as there is no reason to suggest that the level of familiarity or preference for faster voices would differ between male and female voices, or for male and female listeners.

Taken together, the findings suggest that, whilst there is general agreement within the literature that accentuation effects are a real and robust phenomenon with both social and non-social stimuli (e.g., Corneille et al., 2004; Goldstone, 1995; MacLin & Malpass, 2001; Levin & Banaji, 2006; Tajfel & Wilkes, 1963), memory for F0 and speech rate of the voice are unlikely to be affected in this way. This is particularly enlightening given that generalisations in face and voice research have been made, where a finding identified for one is assumed to also be the case for the other (Barsics, 2014). The findings in Experiment 5 (Chapter 8) suggest that differences exist in the mechanisms in memory for both faces and voices, and that researchers should be cautious when making comparisons between them.

11.3.2.1 Summary Conclusions

In summary, Experiment 5a and 5b (Chapter 8) suggests that it is doubtful that listeners rely solely on self-generated categorical information about the voice at the time of encoding to aid recognition of the voice at a later stage. The findings in Experiment 5 (Chapter 8) have therefore contributed further to our understanding of categorisation effects in memory for voices. This work is likely to offer a more robust and informative set of findings than those previously identified in the literature given that a larger set of synthesised male and female voices were used, and manipulations in F0 or speech rate were kept within the ranges that are typical in the English-speaking population. Such work may also prove to be a useful conceptual tool in determining the properties of voice that are more or less affected by intra-individual variation.

11.3.3 Research Question 3: Do listeners become increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition when;

a) F0 is increased and decreased, and the inter-stimulus interval between presentation of the target voice and the sequential voice pair is increased?

b) F0 is increased and decreased, and a different sentence is spoken in the sequential voice pair to the one previously spoken by the target voice?

11.3.3.1 Increasing the Inter-Stimulus Interval

In Experiment 6 (Chapter 9), the inter-stimulus interval between presentation of the target voice and the sequential voice pair was increased from 1-second to 5-seconds. This experiment was designed to push the target voice out of the range, or at least to the very limits, of sensory memory. The acoustic trace is also likely to be weaker when a 5-second inter-stimulus interval is used. Therefore, listeners might become increasingly reliant on category typical representations stored in memory, and we might therefore expect accentuation errors to be made. The pattern of findings in Experiment 6 (Chapter 9) were similar to those observed in Experiment 5a (Chapter 8), when there was a 1-second interval between hearing the target voice and being presented with the sequential voice pair. The results from Experiment 6 (Chapter 9) showed that there was an increase in the selection of voices higher in F0 when both moderate and low F0 target voices were presented. In contrast, no such effect was found for high F0 target voices. Therefore, there was no evidence for accentuation effects for the memory of voice F0 when the interval between hearing the target voice and being asked to recognise this from a voice pair was increased to five seconds. Thus, listeners were no more likely to rely on self-generated categorical information about the voice at the time of encoding to aid recognition when the inter-stimulus interval is increased.

In Experiment 6 (Chapter 9) there were no differences in errors made for high F0 target voices when target voices were paired with distractor voices that were higher and lower in F0. However, in Experiment 5a (Chapter 8) this effect was significant. At present, this finding is difficult to resolve, and may warrant further investigation. Nevertheless, the basic pattern of findings was largely consistent with Experiment 5a (Chapter 8). Thus, it is concluded that listeners were no more likely to rely on self-generated categorical information about the voice at the time of encoding to aid recognition when the inter-stimulus interval is increased.

11.3.3.2 Changing the Spoken Message

In Experiment 7 (Chapter 10), the spoken sentence in the sequential voice pair was different to the one that was previously spoken by the target voice. This experiment was designed to determine whether recognition for the voice would be somewhat more difficult if the sentence spoken was different to the one previously heard. When the same sentence is used, recognition of the voice may be easier because it is achieved by mapping the auditory information of the spoken sentence onto stored representations in memory (Weber & Scharenborg, 2012). By repeating the same sentence, listeners can use patterns identified from the spoken sentence to determine whether the voice heard matches the mental representation stored in memory. Category typical representations stored in memory might also be used more, and we might therefore expect accentuation errors to be made. The pattern of findings in Experiment 7 (Chapter 10) was similar to that observed in both Experiment 5a (Chapter 8) and 6 (Chapter 9). The findings in Experiment 7 (Chapter 10) showed that there was an increase in the selection of distractor voices higher in F0 compared to distractor voices lower in F0 when both moderate and low F0 target voices were presented. In contrast, no such effect was found for high F0 target voices. Therefore, there was no evidence for accentuation effects for the memory of voice F0 when a different sentence is spoken to the one previously heard. Thus, listeners are no more likely to rely on self-generated categorical information about the voice at the time of encoding to aid recognition when a different sentence is spoken to the one previously heard.

One important difference in the findings of Experiment 7 (Chapter 10) compared to those in Experiments 5a (Chapter 8) and 6 (Chapter 9), was that there was an increase in the selection of distractor voices higher in F0 compared to distractor voices lower in F0 for female target voices. In contrast, no difference was found for male target voices. One possible explanation for this finding is that higher F0 female voices may have sounded more like those voices that are typically heard. Indeed, it has been reported that females use upward inflections when they speak more than twice as often as males do (Guy, Horvath, Vonwiller, Daisley, & Rogers, 1986; Hoffman, 2013). Therefore, listeners may be selecting distractor voices higher in F0 compared to distractor voices lower in F0 for female voices because they are familiar with this style of utterance and it sounds like a more plausible version of the target voice (i.e., and inflected version of the target voice). This finding may have only occurred in Experiment 7 (Chapter 10) because listeners are forced to rely on the acoustic properties of the voice (in this case, F0) rather than content when the sentence is different to the one previously heard. Conversely, both Experiment’s 5a (Chapter 8) and 6 (Chapter 9), listeners are able to use similarities in elements of the spoken sentence *and* F0 to help aid the matching process.

Another explanation for this finding is that male voices were actually increased in F0 less than female target voices. A relative percentage change (rather than an absolute percentage change) was used to manipulate the voices in F0. Thus, a percentage change in F0 for male voices was smaller than the same percentage change for female voices because male voices have a lower overall mean F0. It is possible that fewer errors are made for male voices than for female voices because the manipulated versions were more similar in F0 to the target voice. Because listeners are having to rely more on the acoustic properties of the voice (in this case, F0) than on the content of the sentence, they perform better in the matching task for male voices than they do for female voices.

The finding of an increase in the selection of voices higher in F0 than the target voice appears to be reasonably consistent across Experiments 5a (Chapter 8), 6 (Chapter 9), and 7 (Chapter 10) suggesting that the findings are not an anomaly and are robust. Furthermore, it lends support to several conclusions drawn in Experiment 5a (Chapter 8). First, that the accentuation bias is no longer found when a more representative and generalisable set of voices are used, and second, that listeners have difficulty discriminating between voices of higher frequencies. Listeners may have made fewer errors identifying target voices when paired with distractor voices lower in F0 because they were more efficient at detecting the changes in frequency than when distractor voices were higher in frequency. Again, this interpretation would account for why there was no effect of listener sex or sex of voice on errors made identifying target voices. Third, the listeners may be selecting distractor voices higher in F0 more often than distractor voices lower in F0 because they are more familiar with these types of utterances and it sounds like a more plausible version of the target voice (i.e., an inflected version of the target voice). And finally, that the tendency for listeners to select voices higher in F0 compared to voices lower in F0 was strongest for the low F0 target voice condition, because voice higher in F0 are perceived as less threatening or dominant sounding to the listener.

11.3.3.3 Comparison of Recognition Errors Made Overall for Experiments 5a, 6, and 7

Experiment 8 (Chapter 10) investigated whether the amount of matching errors made overall between the three recognition memory experiments (Experiment 5a, Experiment 6, and Experiment 7) were different when manipulations in F0 were made. It was anticipated that listeners might have been reasonably accurate at recognising target voices in Experiment 5a (Chapter 8) because the acoustic trace at short inter-stimulus intervals is likely to be strong. Furthermore, Experiment 5a used the same sentence in the sequential voice pair as the one that was previously spoken by the target voice. Memory for the voice may be better because listeners can use patterns identified from the spoken sentence to determine whether the voice heard matches the mental representation stored in memory. Indeed, recognition of the voice might be accomplished on the basis of a simple familiarity judgement, and without any knowledge of the voice per se, if the same sentence is used.

The findings in Experiment 8 (Chapter 10) indicated that there was no difference in matching errors made overall when a 1-second inter-stimulus interval was used (Experiment 5a, Chapter 8) and when a 5-second inter-stimulus interval was used (Experiment 6, Chapter 9). Recognition tasks are typically easier at shorter retention intervals (i.e., at a 1-second inter-stimulus interval compared to a 5-second inter-stimulus interval) because the acoustic trace is stronger. However, the findings presented in Experiment 8 (Chapter 10) suggest that accurate recognition for the voice does not appear to depend on being able to compare high quality auditory representations in memory. Furthermore, the representation of vocal stimuli stored in auditory sensory memory may be retained for periods of up to 5 seconds, and may leave a stronger trace in memory than non-vocal auditory stimuli. This is inconsistent with existing research that tends to suggest that accuracy for speech stimuli deteriorates quickly, and over a matter of seconds (e.g., Crowder, 1982; Hanson, 1977; Pisoni, 1973). Rather the findings in Experiment 8 (Chapter 10) offer support to others who have found that increasing the length of the inter-stimulus interval did not affect matching accuracy for the voice (Smith, Dunn, Baguley, and Stacey, 2016). The findings in Experiment 8 (Chapter 10) also suggest that memory for F0 of the voice may be retained for periods for up to five seconds, and for periods longer than when the acoustic trace has supposedly degraded and is weak (Glanzer & Cunitz, 1966; Lu, Williamson & Kaufman, 1992; Treisman, 1964; Wickelgren, 1969).

Listeners made fewer matching errors overall when the content of the sentence in the sequential voice pair was the same as the sentence spoken by the target voice (Experiment 5a, Chapter 8) compared to when the content of the sentence in the sequential voice pair was different to the sentence spoken by the target voice (Experiment 7, Chapter 10). This finding offers support to others who have found recognition tasks to be typically harder when a different sentence is spoken to the one previously heard (e.g. Gliskey et al., 2001; Reid & Craik, 1995; Winograd, Kerr, & Spence, 1984). This suggests that listeners might be relying on elements of the spoken sentence to help aid the recognition process. In Experiment 7 (Chapter 10), listeners were forced to rely on acoustic properties of the voice (in this case, F0) rather than on elements of the spoken sentence, because a different sentence was used. Conversely, in Experiment 5a (Chapter 8), listeners were able to use elements of the spoken sentence *and* F0 to help aid recognition, making the task easier for listeners.

11.3.3.4 Summary Conclusions

In summary, Experiments 5a, 6, and 7 suggest that listeners are susceptible to distortions in memory for F0. However, the findings were difficult to explain using the accentuation effect. Based on the findings here, it is doubtful that listeners rely on self-generated categorical information about the voice at the time of encoding to aid recognition when a 1-second inter-stimulus interval is used between presentation of the target voice and the sequential voice pair. Furthermore, it is unlikely that listeners become increasingly reliant on self-generated categorical information about the voice at the time of encoding to aid recognition when the inter-stimulus interval between presentation of the target voice and the sequential voice pair is increased to 5-seconds, or when a different sentence is spoken in the sequential voice pair to the one that was previously spoken by the target voice. Listeners made fewer matching errors overall when the same sentence, compared to when a different sentence is used in the sequential voice pair. Therefore, it is likely that listeners use both elements of the spoken sentence and F0 to help aid recognition. However, there was no difference in errors between the 1- and 5-second inter-stimulus interval condition. Thus, it appears that the representation of vocal stimuli stored in auditory sensory memory may be retained for periods of up to 5 seconds, and may leave a stronger trace in memory than non-vocal auditory stimuli.

11.4 Applied Implications of the Research Findings

There are several criminal situations in which voices are the most distinct and reliable cue to a speaker’s personal characteristics, including when visual conditions are poor, when the face of a target is covered, or when a crime is committed over the phone (Yarmey et al., 1996; Yarmey, 2001, 2004; Waller & Eriksson, 2016). In such cases, descriptions of the perpetrator may be based solely on information from the voice. Following a crime, the police might ask the victim and/or witness to provide characteristic information about the voice of a suspect (Waller & Eriksson, 2016). This earwitness information might be useful to the police in narrowing down a list of suspects. However, this might pose a problem if descriptions made about the characteristics of the voice by the victim and/or witness are inaccurate.

The findings from Experiments 2, 3, and 4 suggest that listeners do use acoustic cues of the voice when making judgements about the characteristics of the speaker. However, it is important to emphasise that not all acoustic cues are likely to affect perceptual judgements about the characteristics of the speaker. Manipulations in F0 are more disruptive to perceptions of speaker identity and sex than manipulations in speech rate. However, for speaker sex, the effect of these manipulations in F0 appear to be markedly different for male and female voices. In contrast, manipulations in both F0 and speech rate are likely to affect perceptions of speaker age. Therefore, a victim and/or witness’s perception of the suspect’s personal characteristics (or at least for the suspects identity, sex, and age) are likely to be influenced by changes made by the speaker in the F0 or speech rate of their voice. These changes might be particularly effective at changing the identity, sex, and age of a perpetrator if they are intentionally trying to modify their own voice through means of disguise. It is important for law enforcers to have an understanding about what cues of the voice are perceptually important and which ones do not produce any noticeable changes in a person’s voice. A more in depth understating of this information is likely to help determine the accuracy of descriptions made about a voice, particularly if a suspect of a crime was likely to be disguising their voice when the crime took place.

Following a crime, the police may also decide to conduct a voice lineup, which requires the victim and/or witnesses to identify a suspect from their voice. Such evidence can be admitted to court, and often constitutes as pervasive and pivotal evidence in a case (Overbeck, 2002). However, inaccurate identification may lead to the prosecution of innocent persons while the guilty party goes free. It is therefore important for the legal system to have knowledge about the role of acoustic cues of the voice, how manipulations in these can affect recognition performance, and the pattern of errors that might arise because of these.

The findings from Experiment 5a, 5b, 6, and 7 suggest that manipulations in F0 or speech rate do affect recognition performance for the voice. Although recognition performance was above chance (i.e., listeners were more likely to correctly identify the target voice than incorrectly identify a manipulated version of the target voice), there are considerable consequences involved when any error is made. Indeed, the innocent may be punished for a crime they did not commit, or the perpetrator may be incorrectly released. A more in depth understating of this information is likely to be insightful to the police and help to determine the accuracy of decisions made during a voice lineup, particularly if a suspect of a crime was likely to be disguising their voice

The findings in Experiment 8 (Chapter 10) may also be useful when determining the most appropriate method of conducting a lineup. Several experts have suggested criteria for voice lineups based on the findings of their work (Broeders & Rietveld, 1995; Bull & Clifford, 1999; Hammersley & Read, 1996; Hollien, 1996, 2002; Hollien, Huntley, Kunzel & Hollien, 1995; Ormerod, 2001). One suggestion includes that the lineup should not contain words or phrases spoken by the perpetrator during the time of the crime to prevent deliberate distortion of specific words or phrases by a guilty suspect at the time of recording during the investigation stage. Nevertheless, the findings in Experiment 8 (Chapter 10) suggest that different words or phrases spoken by the suspect to those heard at the time of the crime might introduce further errors and reduce recognition performance for the voice even more. Rather, it might be more appropriate for the same phrases to be spoken by the suspect at the time of the crime to aid recognition.

11.5 Limitations of the Thesis

Experiments 2, 3, 4, 5a, 5b, and 6 in this thesis used the same sentence throughout (e.g., *"Spring is the season where flowers appear, summer is the warmest season of the year"*). It was important to control for content of the spoken sentence so that findings could confidently be compared with each other, both within the same experiment, and across different experiments. The aim of this thesis was to determine the effect of manipulations in F0 or speech rate on perceptions about the speaker and voice recognition performance. The use of different sentences, either within the same experiment or across different experiments, would have made it difficult to determine whether the findings were attributable to the manipulation in F0 or speech rate, or whether they were due to changes to the sentence. Arguably, there might be something special about this sentence and findings in this thesis might apply to this particular sentence. However, Experiment 7 used a different sentence spoken in the sequential voice pair to the target voice previously heard and the pattern of findings were similar to those observed in both Experiment 5a and 6. Therefore, it seems unlikely that the use of different sentences would have altered the findings and conclusions made.

Experiments 5a, 5b, 6, and 7 in this thesis each used a sample size of thirty participants. For the purposes of this thesis, this was deemed appropriate as these experiments followed the methodology and sample sizes employed by previously published work (e.g., Mullenix et al., 2010; Stern et al., 2007). However, it should be noted that the analytical design employed the use of a between subjects variable of listener sex (i.e., male or female). Consequently, this meant that participant numbers were lower (i.e., 15 males and 15 females) for several conditions in the analysis. It seems unlikely that this would have had an impact on the overall outcome of the analysis, nevertheless, it is possible that some smaller effects may have been undetected.

The stimuli used in this thesis were synthesised voices. Synthetic speech was used because of the need for precisely controlled stimuli that varied in F0 or speech rate and to ensure that the voices used were unfamiliar to the listeners. Synthetic speech was also used because the some of the experiments carried out were following the methodology used in previously published work (e.g., Mullenix et al., 2010; Stern et al., 2007). Natural Reader 12.0 was chosen to obtain the synthesised speech samples because it generates speech form concatenated pieces of real human speech that are realistic and natural sounding. Indeed, mean naturalness ratings averaged above 70% for the original voices and their manipulated versions, and are a good indication that the synthesised voices used for the experiment were representative of real voices (refer to Chapter 4, Section 4.3.3 and 4.3.4). These values are also similar to those identified in the literature (Jreige et al., 2009).

The experiments in this thesis also used a more representative and generalisable set of voices than in previous experiments on speaker perception and recognition memory. However, it is important to note that some of the voices may have sounded less realistic than others. For example, in natural speech, a person speaking more slowly is likely to be more hesitant, making more silent pauses or filled pauses (e.g., *um,* er). In this thesis, decreasing speech rate did affect the rate of continuous production but did not lead to increased pauses of any kind. Therefore, the speech samples used might not have been an entirely natural rendition of slower speech, at least of a type that listeners most typically hear. In Experiment 5b, it is possible that at the lower margins of the speech rate manipulated samples (i.e., the slowest samples), but not elsewhere, the participants may have selected a faster voice in the pair because it sounded more realistic. Furthermore, the naturalness ratings for the voices with higher F0 manipulations in Experiments 5a, 6, and 7 yielded slightly higher naturalness rating scores than those with lower manipulations. It is possible that the listeners may have preferred the more natural sounding voices (i.e., the higher F0 manipulations) and were thus, more likely to select them. Nevertheless, given that the voices were generally perceived to be natural sounding across the board, with differences observed between the voices being relatively small, it is unlikely that this would have impacted significantly upon the matching tasks.

Relative, rather than absolute, percentage changes were used to manipulate the voices in this thesis. Relative percentage change takes into account the overall frequency, or speech rate, of the stimuli being manipulated. Thus, each voice is manipulated by the same percentage in relation to the mean F0 or speech rate. It was felt that this was appropriate to ensure that all manipulations made were proportionally the same across the voices used, and so that findings could be compared with each other. However, this meant that when F0 was increased for the male voices, the manipulated versions did not fall in either the gender ambiguous range or the typical female F0 range for voiced speech. Therefore, in Experiment 3 (Chapter 6), it is difficult to determine whether male voices that fell within the typical female F0 range for voiced speech would have increased the likelihood that these would be perceived as female. Nevertheless, given that the findings in Experiment 3 (Chapter 6) showed that increasing F0 did not change the perceived sex of the speaker for male voices at all, it is unlikely that listeners would have perceived the voices as female, or at least at a greater percentage of the time than they would have perceived the voices as male.

The actual ages of the voices used in this thesis were unknown. This was not important in order to investigate the main aims of the thesis. Indeed, the author was interested in whether manipulations in F0 or speech rate affect perceptions of certain characteristics of the speaker. However, this knowledge would have been useful to determine the actual age difference observed between the original voices and their manipulated versions in Experiment 4 (Chapter 7). This would be particularly useful information for the police during a criminal investigation to help to determine possible age parameters for a suspect, especially if they were concerned that voice disguise may have been used when the crime took place.

One final limitation of the stimuli set concerns the use of a different sentence in Experiment 7 (Chapter 10). The different sentence that was used in this experiment (i.e., sentence two) did not undergo the same rigorous testing and validation as the first sentence that was used in Experiments 2, 3, 4, 5a, 5b, 6, and 7. Rather, inferences were made about the effect of the voice manipulations and how natural the voices sounded from the findings using sentence one. This was deemed to be sufficient because it was not expected that any significant differences would have been observed between the two sentences used. Nevertheless, it would be worth checking this to ensure that this is the case and that there is indeed consistency across the findings when different sentences are used and manipulations in F0 or speech rate are made.

11.6 Outstanding Research Questions and Possible Future Directions

The findings in this thesis offer several recommendations for future research. The most specific recommendation relates to the voice stimuli that were used. As previously noted, synthesised voices were used because of the need for precisely controlled stimuli. Furthermore, much of the existing work carried out so far has used synthesised voices (e.g., Mullenix et al., 2010; Stern et al., 2007), and it was important for the experiments to uphold some of the key features of the work in the existing literature. The use of synthesised voices does not detract away from the experiments carried out in this thesis and the conclusions that have been made. Indeed, Experiment 1c and 1d (Chapter 4, Section 4.4) identified that listeners did believe that the voices sounded natural and realistic. Therefore, the voices used in this thesis are likely to be generalisable to those voices that are heard in the real-world. As such, it is probable that findings with real voices would be similar to those that have been identified in this thesis. Nevertheless, it is necessary to acknowledge the use of real voices as a potential avenue for work to be carried out in the future.

Future work should also consider the effect of manipulations in other acoustic cues of the voice and determine their effect on perceptual judgements about characteristics of the speaker and recognition performance for the voice. Owing to the time constraints in producing a PhD thesis, it would have been difficult to study the effect of all cues here. Furthermore, a thorough amount of time was spent developing the stimuli to ensure strong grounds on which to make any claims. It was also important to address those cues that had previously been investigated as there were several outstanding and unanswered questions that the author wanted to resolve. The findings of this thesis suggest that the likelihood that a particular acoustic cue will affect recognition performance and perceptions of the speaker are dependent on the acoustic cue under investigation. Indeed, F0 appears to be the more useful cue in voice matching tasks and when determining certain characteristics of the speaker. If this pattern of findings holds true for other properties of speech too, a dichotomy between those properties that are more, or less, susceptible to manipulations might be made. This could prove to be a useful conceptual tool in categorising various attributes that compose a speaker’s voice, and help to predict the likelihood of errors occurring when manipulations are made. This would be particularly useful when descriptions about the voice are given by a victim and/or witness, and when voice lineups are used as part of a criminal investigation.

A further recommendation involves the use of voices of different nationalities. In this thesis, English participants were tested with exclusively English stimuli, all of which had a similar regional accent. It was important to control for this in the experiments given that regional accent has been found to affect recognition performance for voices. Indeed, studies on own- and other-race/ethnicity in voices have shown that people may be better at recognising voices of their own race/ethnicity than those of another race/ethnicity (e.g., Doty, 1998; Goggin, Thompson, Stevenage, Clarke & McNeill, 2012; Strube, & Simental, 1991; Hollien, Majewski, & Doherty, 1982; Koster & Schiller, 1997; Koster, Schiller, Kunzel, 1995; Schiller & Koster, 1996). Similarly, the other-accent effect suggests that listeners are better able to recognise speakers with a more familiar, or a more similar, accent (Goldstein, Knight, Bailis, & Conover, 1981; Thompson, 1987; Vanags, Carrol, & Perfect, 2005). In keeping with the own race/ethnicity and accent bias, the ability to accurately recognise voices might have an important cultural underpinning relating to expertise and exposure (Levin, 2000; Meissner & Brigham, 2001; Tanaka, 2001). Furthermore, expertise and exposure might play a role in enabling accurate recognition and perceptual judgements about the speaker because of cultural differences in voice production. Indeed, research has shown that speakers of different languages or dialects may use characteristically different ranges and typical F0 values (Dolson, 1994). For example, Japanese females have been found to exhibit a higher mean F0 than American (Yamazawa & Hollien, 1992) and English (Loveday, 1981; van Bezooijen, 1995; 1996; Yamazawa & Hollien, 1992) speakers. Additionally, Japanese females tend to produce a bimodal F0 distribution pattern, whereas American and English speakers tend to produce a unimodal distribution (Yamazawa & Hollien, 1992). The bimodal distribution is explained by the high-low tone contrast present in the Japanese language (Heffernan, 2007). This could make it difficult for people to match voices speaking a different language.

Future work could also consider the role of individual differences in participants' abilities in recognising voices. Research has shown that face recognition skills are subject to wide individual variation, with some people showing exceptional ability – a group that has come to be known as 'super-recognisers' (Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016). It may be that some individuals are particularly good at recognising voices. Future work could therefore aim to establish with this is indeed the case. Furthermore, if there were 'super-recognisers' for voices, research could look to investigate whether such individuals are any less susceptible to people who have disguised their voice by making changes to certain acoustic properties (e.g., F0 or speech rate). Such findings would have real world implications, particularly with the police in forensic and security operations. Indeed, the Metropolitan Police Force in London already recruits 'super-recognisers within its ranks for deployment on various identification tasks.

It is important to be careful when generalising the results of laboratory experiments to real-world situations. In terms of earwitness testimony, when a person is involved in a real-world incident, there are likely to be other factors contributing to the accuracy of speaker memory. For instance, it is likely that a higher stress level is present in the earwitness situation, due to personal threat or heightened personal arousal (Wilding et al., 2000). It is possible that cognitive processes producing the errors in memory that the author has described are susceptible to stress factors. Future work could therefore examine more stress induced and emotionally arousing scenarios, such as the weapon focus effect, that presumably tap into stress reactions that may be useful in examining the degree to which errors in memory occur (Loftus, Liftus, & Messo, 1987).

In the interest of ecological validity, it is also important to examine variables that are often encountered during the course of a criminal investigation. One notable variable is the retention interval between when an event is witnessed and when a victim and/or a witness is asked to provide details about the suspect, or identify the suspect from a voice lineup. In a real-world criminal situation, there is uncertainty over the time period between hearing a voice and being asked to identify the voice at a later date. Furthermore, studies have typically shown that longer retention intervals reduce recognition performance for voices (e.g., Kerstholt et al., 2004, 2006; van Wallendael, Surace, Parson, & Brown, 1994; Yarmey & Matthys, 1992). Future work should therefore consider longer retention intervals, such as days and weeks, between presentation of the target voice and being asked to recognise this at a later date.

One final recommendation involves the use of different speakers being used in a lineup style procedure. This thesis has considered whether manipulations in F0 or speech rate affect recognition performance for a voice using 2AFC tasks where the original version of a voice is paired with a manipulated version of a voice. It would be interesting to determine whether listeners are able to correctly identify the original version of a target voice that is then manipulated in either F0 or speech rate and presented with several different speakers (i.e., different identities) rather than the same speaker. This procedure would be more similar to a situation during a criminal investigation, such as when a suspect has disguised their voice (i.e., by manipulating it in either F0 or speech rate) when a voice lineup is conducted and different people are used as fillers.

11.7 Concluding Comments

The central aim of this thesis was to determine whether manipulations in F0 or speech rate affect perceptions about several characteristics of the speaker. The work also set out to determine the effect of manipulations in F0 or speech rate on recognition performance for the voice. The findings presented suggest that, at least for unfamiliar synthesised voices, manipulations in F0 are likely to affect perceptions of the identity and age of the speaker. For female voices, decreasing F0 also increased the likelihood that the voices would be perceived as male. Manipulations in speech rate are unlikely to change perceptions of the identity or the sex of the speaker. However, manipulations in speech rate do appear to affect perceptions of speaker age. Thus, the likelihood that a particular acoustic cue will affect perceptual judgements about certain characteristics of the speaker is likely to be dependent both on the characteristic and the acoustic cue under investigation.

The findings have also shown that listeners are susceptible to making errors in memory for voices when manipulations in F0 or speech rate are made. However, the findings are difficult to explain using the accentuation effect. It is therefore doubtful that listeners rely solely on self-generated categorical information about the voice at the time of encoding to aid recognition. Furthermore, for F0, listeners are no more likely to rely on self-generated categorical information about the voice at the time of encoding to aid recognition when the inter-stimulus interval is increased, or when a different sentence is spoken in the sequential voice pair to the one that was previously spoken by the target voice. Listeners were found to make fewer matching errors overall when the same sentence, rather than a different sentence, was used in the sequential voice pair as the previously heard target voice. Therefore, it is likely that listeners rely on elements of the spoken sentence to help aid the recognition process. However, there was no difference in errors made overall when a 1-second inter-stimulus interval was used and when a 5-second inter-stimulus interval was used between presentation of the target voice and the sequential voice pair. Thus, it appears that accurate recognition for a voice does not depend on being able to compare high quality auditory representations in memory. Furthermore, the representation of vocal stimuli stored in auditory sensory memory may be retained for periods of up to 5 seconds, and may leave a stronger trace in memory than non-vocal auditory stimuli.

This thesis has made an independent contribution to knowledge and advanced on the research currently being carried out in this domain, by determining the properties of voice that are more or less affected by intra-individual variation, whether it be through unintentional or deliberate means. The work has provided a more detailed understanding of the acoustic properties of the voice that are likely to affect perceptual judgements about the identity, sex, and age of the speaker than has been possible to establish from previous studies, given several methodological issues with the work carried out so far. The findings have also contributed further to our understanding about the impact of manipulations of F0 or speech rate on recognition performance, and the mechanisms important for accurate voice recognition. In light of the applied relevance of the findings, this topic is undoubtedly an important one that should continue to be explored further.

REFERENCES

Abberton, E. R., Howard, D. M., & Fourcin, A. J. (1989). Laryngographic assessment of normal voice: A tutorial. *Clinical Linguistics & Phonetics, 3*(3), 281-296.

Abberton, E., & Fourcin, A. J. (1978). Intonation and speaker identification. *Language and Speech, 21*(4), 305-318.

Abbs, J. H. (1973). The influence of the gamma motor system on jaw movements during speech: A theoretical framework and some preliminary observations. *Journal of Speech, Language, and Hearing Research, 16*(2), 175-200.

Adams, J. A. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin, 101*(1), 41.

Adams, S. G., Weismer, G., & Kent, R. D. (1993). Speaking rate and speech movement velocity profiles. *Journal of Speech, Language, and Hearing Research, 36*(1), 41-54.

Amino, K., & Arai, T. (2009). Effects of linguistic contents on perceptual speaker identification: Comparison of familiar and unknown speaker identifications. *Acoustical Science and Technology, 30*(2), 89-99.

Anderson, T. D., & Sataloff, R. T. (2004). Complications of collagen injection of the vocal fold: Report of several unusual cases and review of the literature. *Journal of Voice, 18*(3), 392-397.

Andrews, M. L., & Schmidt, C. P. (1997). Gender presentation: Perceptual and acoustical analysesof voice. *Journal of Voice, 11*(3), 307-313.

Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology, 37*(5), 715.

Ardran, G., & Kemp, F. (1966). The mechanism of the larynx part I: The movements of the arytenoid and cricoid cartilages. *The British Journal of Radiology, 39*(465), 641-654.

Arnfield, S., Roach, P., Setter, J., Greasley, P., & Horton, D. (1995). Emotional stress and speech tempo variation. *Speech Under Stress,* 13-15.

Aronovitch, C. D. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *The Journal of Social Psychology, 99*(2), 207-220.

Aronson, A. E., & Bless, D. (2011). *Clinical voice disorders.* Thieme.

Assmann, P. F., Dembling, S., & Nearey, T. M. (2006). Effects of frequency shifts on perceived naturalness and gender information in speech. *Interspeech,* 889-892.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation, 2,* 89-195.

Bachorowski, J. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science, 8*(2), 53-57.

Bachorowski, J., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The Journal of the Acoustical Society of America, 106*(2), 1054-1063.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8,* 47-89.

Baken, R., & Orlikoff, R. (1987). Clinical measures of speech and voice. *Boston: College-Hill,*

Baldwin, C. L. (2012). *Auditory cognition and human performance: Research and applications* CRC Press.

Barlow, S. M., Finan, D. S., Bradford, P. T., & Andreatta, R. D. (1993). Transitional properties of the mechanically evoked perioral reflex from infancy through adulthood. *Brain Research, 623*(2), 181-188.

Barsics, C. (2014). Person recognition is easier from faces than from voices. *Psychologica Belgica, 54*(3), 244-254.

Basow, S. A. (1992). *Gender: Stereotypes and roles* Wadsworth Publishing Company.

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology, 102*(4), 711-725.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129-135.

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research, 13*(1), 17-26.

Belizaire, G., Fillion-Bilodeau, S., Chartrand, J. P., Bertrand-Gauvin, C., & Belin, P. (2007). Cerebral response to 'voiceness': A functional magnetic resonance imaging study. *Neuroreport, 18*(1), 29-33. doi:10.1097/WNR.0b013e3280122718 [doi]

Benavides, A. M., Murillo, J. L. B., Pozo, R. F., Cuadros, F. E., Toledano, D. T., Alcázar-Ramírez, J. D., & Gómez, L. A. H. (2016). Formant frequencies and bandwidths in relation to clinical variables in an obstructive sleep apnea population. *Journal of Voice, 30*(1), 21-29.

Benguérel, A., & Cowan, H. A. (1974). Coarticulation of upper lip protrusion in french. *Phonetica, 30*(1), 41-55.

Benjamin, B. J. (1981). Frequency variability in the aged voice. *Journal of Gerontology, 36*(6), 722-726.

Bennett, S., & Montero-Diaz, L. (1982). Children's perception of speaker sex. *Journal of Phonetics, 10*(1), 113-121.

Benninger, M. S., & Murry, T. (2008). *The singer's voice.* Plural Publishing.

Berry, J. (2002). Acoustic effects of speaking rate changes in articulatory models. *The Journal of the Acoustical Society of America, 112*(5), 2442-2442.

Block, S., & Killen, D. (1996). Speech rates of Australian English-speaking children and adults. *Australian Journal of Human Communication Disorders, 24*(1), 39-44.

Boersma, P., & Weenink, D. (2006). Praat (version 4.5) [computer software]. *Amsterdam: Institute of Phonetic Sciences.*

Bolinger, D. (1964). Intonation as a universal. *Proceedings of the 5th Congress of Phonetics, Cambridge 1962,* 833-848.

Bond, R. N., & Feldstein, S. (1982). Acoustical correlates of the perception of speech rate: An experimental investigation. *Journal of Psycholinguistic Research, 11*(6), 539-557.

Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour, 82*(1), 55-59.

Boyce, S. E., & Krakow, R. A., Bell-Berti, F., & Gelfer, C. E. (1990). Converging sources of evidence for dissecting articulatory movements into core gestures. *Journal of Phonetics, 18*, 173-188.

Braun, A., Rietveld, T., & van Bezooijen, R. (1995). The influence of smoking habits on perceived age. *Proceedings of the XIIIth International Congress of Phonetic Sciences, 2,* 294-297.

Bregman, M., & Creel, S. (2012). Learning to recognize unfamiliar voices: the role of language familiarity and music experience. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34, No. 34).

Brizendine, L. (2006). *The female brain.* Broadway Books. Morgan Road, New York.

Broeders, A., & Rietveld, A. (1995). Speaker identification by earwitnesses. *Beiträge Zur Phonetik Und Linguistik, 64*, 24-40.

Broeders, A., & Rietveld, A. (1995). Speaker identification by earwitnesses. *Beiträge Zur Phonetik Und Linguistik, 64*, 24-40.

Brosch, T., Pourtois, G., & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion, 24*(3), 377-400.

Brown, B. L., Strong, W. J., & Rencher, A. C. (1974). Fifty-four voices from two: The effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech. *The Journal of the Acoustical Society of America, 55*(2), 313-318.

Brown, R. (1981). An experimental study of the relative importance of acoustic parameters for auditory speaker recognition. *Language and Speech, 24*(4), 295-310.

Brückl, M., & Sendlmeier, W. (2003). Aging female voices: An acoustic and perceptive analysis. *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis,* 163-168.

Buhusi, C. V., & Meck, W. H. (2005). What makes us tick? functional and neural mechanisms of interval timing. *Nature Reviews. Neuroscience, 6*(10), 755.

Bull, R., & Clifford, B. R. (1984). Earwitness voice recognition accuracy. *Eyewitness Testimony: Psychological Perspectives,* 92-123.

Bull, R., & Clifford, B. R. (1999). Earwitness testimony. *Medicine, Science and the Law, 39*(2), 120-127.

Buller, D. B., & Aune, R. K. (1988). The effects of vocalics and nonverbal sensitivity on compliance a speech accommodation theory explanation. *Human Communication Research, 14*(3), 301-332.

Buller, D. B., & Burgoon, J. K. (1986). The effects of vocalics and nonverbal sensitivity on compliance A replication and extension. *Human Communication Research, 13*(1), 126-144.

Busby, P. A., & Plant, G. (1995). Formant frequency values of vowels produced by preadolescent boys and girls. *The Journal of the Acoustical Society of America*, *97*(4), 2603-2606.

Byrd, D. (1992). Preliminary results on speaker-dependent variation in the TIMIT database. *The Journal of the Acoustical Society of America, 92*(1), 593-596.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication, 15*(1-2), 39-54.

Byrd, D., & Tan, C. C. (1996). Saying consonant clusters quickly. *Journal of Phonetics, 24*(2), 263-282.

Callan, D. E., Kent, R. D., Guenther, F. H., & Vorperian, H. K. (2000). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language, and Hearing Research, 43*(3), 721-736.

Cavanagh, S. E. (2011). Early pubertal timing and the union formation behaviors of young women. *Social Forces, 89*(4), 1217-1238.

Chatterjee, I., Halder, H., Bari, S., Kumar, S., Roychoudhury, A., & Murthy, P. (2011). An analytical study of age and gender effects on voice range profile in Bengali adult speakers using phonetogram. *International Journal of Phonosurgery and Laryngology, 1*(2), 65-70.

Chatterjee, K., Tuli, S., Pickering, S. G., & Almond, D. P. (2011). A comparison of the pulsed, lock-in and frequency modulated thermography nondestructive evaluation techniques. *NDT & E International, 44*(7), 655-667.

Chhetri, D. K., Neubauer, J., & Berry, D. A. (2012). Neuromuscular control of fundamental frequency and glottal posture at phonation onset. *The Journal of the Acoustical Society of America, 131*(2), 1401-1412.

Ching, M. K. (1982). The question intonation in assertions. *American Speech,* 95-107.

Choi, J., Hasegawa-Johnson, M., & Cole, J. (2005). Finding intonational boundaries using acoustic cues related to the voice source. *The Journal of the Acoustical Society of America, 118*(4), 2579-2587.

Chon, H. C., Ko, D. H., & Shin, M. J. (2004). Disfluency characteristics and speech rate of stuttering and nonstuttering. *Communication Sciences & Disorders, 9*(2), 102-115.

Clark, J., & Yallop, C. (1995). An introduction to phonetics and phonology. *Blackwell Publishing, London.*

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior, 4*(4), 373.

Clifford, B. R., Rathborn, H., & Bull, R. (1981). The effects of delay on voice recognition accuracy. *Law and Human Behavior, 5*(2-3), 201-208.

Clifford, J. (1983). On ethnographic authority. *Representations,* 2, 118-146.

Cobb, N. J., Lawrence, D. M., & Nelson, N. D. (1979). Report on blind subjects tactile and auditory recognition for environmental stimuli. *Perceptual and Motor Skills, 48*(2), 363-366.

Coleman, R. O. (1976). A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech and Hearing Research, 19*(1), 168-180.

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour, 60*(6), 773-780.

Cook, S., & Wilding, J. (1997). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology, 11*(2), 95-111.

Corneille, O., Huart, J., Becquart, E., & Brédart, S. (2004). When memory shifts toward more typical category exemplars: Accentuation effects in the recollection of ethnically ambiguous faces. *Journal of Personality and Social Psychology, 86*(2), 236.

Corr, P. J. (2013). Approach and avoidance behaviour: Multiple systems and their interactions. *Emotion Review, 5*(3), 285-290.

Crowder, R. G. (1982). The demise of short-term memory. *Acta Psychologica, 50*(3), 291-323.

Cruttenden, A. (1986). 1997. *Intonation.* Cambridge University Press, Australia.

Crystal, D. (2006). How language works: How babies babble, words change meaning, and language lives or dies.

Crystal, T., & House, A. (1986). Characterization and modelling of speech-segment durations. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86. 11,* 2791-2794.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *The Journal of the Acoustical Society of America, 83*(4), 1553-1573.

Cumming, R. (2011). The effect of dynamic fundamental frequency on the perception of duration. *Journal of Phonetics*, *39*(3), 375-387.

Davenport, M., Davenport, M., & Hannahs, S. (2010). *Introducing phonetics and phonology* Routledge, Oxon.

Dehqan, A., Scherer, R. C., Dashti, G., Ansari-Moghaddam, A., & Fanaie, S. (2012). The effects of aging on acoustic parameters of voice. *Folia Phoniatrica Et Logopaedica : Official Organ of the International Association of Logopedics and Phoniatrics (IALP), 64*(6), 265-270.

D'haeseleer, E., Van Lierde, K., Claeys, S., & Depypere, H. (2012). The impact of menopause and hormone therapy on voice and nasal resonance. *Facts, Views & Vision in ObGyn, 4*(1), 38-41.

Diehl, R. L., Lindblom, B., Hoemeke, K. A., & Fahey, R. P. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics, 24*(2), 187-208.

Diehl, R. L. (2008). Acoustic and auditory phonetics: The adaptive design of speech sound systems. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 363*(1493), 965-978.

Dlugan, A. (November, 2012). *What is the Average Speaking Rate?* Retrieved from http://sixminutes.dlugan.com/speaking-rate/.

Dolson, M. (1994). The pitch of speech as a function of linguistic community. *Music Perception: An Interdisciplinary Journal*, *11*(3), 321-331.

Doty, N. D. (1998). The influence of nationality on the accuracy of face and voice recognition. *The American Journal of Psychology, 111*(2), 191.

Endres, W., Bambach, W., & Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. *The Journal of the Acoustical Society of America, 49*(6B), 1842-1848.

Engstrand, O. (1988). Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *The Journal of the Acoustical Society of America, 83*(5), 1863-1875.

Ericsdotter, C., & Ericsson, A. M. (2009). Gender differences in vowel duration in read Swedish: Preliminary results. *Working Papers in Linguistics, 49,* 34-37.

Evans, S., Neave, N., & Wakelin, D. (2006). Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology, 72*(2), 160-163.

Eysenck, M. W., & Keane, M. T. (2000). *Cognitive psychology: A student's handbook.* Taylor & Francis, East Sussex, England.

Fairbanks, G. (1940). Recent experimental investigations of vocal pitch in speech. *The Journal of the Acoustical Society of America, 11*(4), 457-466.

Fairbanks, G., & Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion∗. *Communications Monographs, 6*(1), 87-104.

Fant, G. (1960). *Acoustic theory of speech perception.* Mouton & Company, The Netherlands.

Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report, 1*, 22-30.

Fant, G. (1970). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations.* Mouton & Company, The Netherlands.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour, 69*(3), 561-568.

Ferrand, C. T. (2002). Harmonics-to-noise ratio: An index of vocal aging. *Journal of Voice, 16*(4), 480-487.

Fiske, S. T., Gilbert, D. T., & Lindzey, G. (2010). *Handbook of social psychology.* John Wiley & Sons, United Kingdom.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 106*(3), 1511-1522.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 106*(3), 1511-1522.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 106*(3), 1511-1522.

Fitch, W. T., & Reby, D. (2001). The descended larynx is not uniquely human. *Proceedings Biological Sciences, 268*(1477), 1669-1675.

Fitzsimons, M., Sheahan, N., & Staunton, H. (2001). Gender and the integration of acoustic dimensions of prosody: Implications for clinical studies. *Brain and Language, 78*(1), 94-108.

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *The Journal of the Acoustical Society of America, 84*(1), 70-79.

Flege, J. E. (1988). The production and perception of foreign language speech sounds. *Human Communication and its Disorders: A Review, 2*, 224-401.

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., & Liss, J. M. (2015). The relationship between speech segment duration and vowel centralization in a group of older speakers. *The Journal of the Acoustical Society of America, 138*(4), 2132-2139.

Fougeron, C., Kuehnert, B., Imperio, M., & Vallee, N. (2010). *Laboratory phonology.* Mouton de Gruyter, Berlin.

Foulkes, P., & Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics, 7*, 180-198.

Fraccaro, P. J., O'Connor, J. J., Re, D. E., Jones, B. C., DeBruine, L. M., & Feinberg, D. R. (2013). Faking it: Deliberately altered voice pitch and vocal attractiveness. *Animal Behaviour, 85*(1), 127-136.

Garnier, M., Wolfe, J., Henrich, N., & Smith, J. (2008, January). Interrelationship between vocal effort and vocal tract acoustics: A pilot study. *Ninth Annual Conference of the International Speech Communication Association,* Gaudrain, E., Li, S., Ban, V. S., & Patterson, R. D. (2009). *The role of glottal pulse rate and vocal tract length in the perception of speaker identity.* Paper presented at *Tenth Annual Conference of the International Speech Communication Association,* Brisbane, Australia.

Gay, T., & Hirose, H. (1973). Effect of speaking rate on labial consonant production. *Phonetica, 27*(1), 44-56.

Gelfer, M. P. (1993). A multidimensional scaling study of voice quality in females. *Phonetica, 50*(1), 15-27.

Gelfer, M. P., & Bennett, Q. E. (2013). Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender. *Journal of Voice, 27*(5), 556-566.

Gelfer, M. P., & Mikos, V. A. (2005). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of Voice, 19*(4), 544-554.

Gelfer, M. P., & Schofield, K. J. (2000). Comparison of acoustic and perceptual measures of voice in male-to-female transsexuals perceived as female versus those perceived as male. *Journal of Voice, 14*(1), 22-33.

Gescheider, G. (1997). *Chapter 3. the classical psychophysical methods. Psychophysics: The fundamentals* (3rd Ed). Lawrence Erlbaum Associates.

Ghazanfar, A. A., & Rendall, D. (2008). Evolution of human vocal production. *Current Biology, 18*(11), R457-R460.

Gifford, A. M., Cohen, Y. E., & Stocker, A. A. (2014). Characterizing the impact of category uncertainty on human auditory categorization behavior. *PLoS Computational Biology, 10*(7), e1003715.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5*(4), 351-360.

Glisky, E. L., Rubin, S. R., & Davidson, P. S. (2001). Source memory in older adults: An encoding or retrieval problem? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1131.

Goffman, L., Maassen, B., & Van Lieshout, P. (2010). Dynamic interaction of motor and language factors in normal and disordered development. *Speech Motor Control: New Developments in Basic and Applied Research,* 137-152.

Goffman, L., & Smith, A. (1999). Development and phonetic differentiation of speech movement patterns. *Journal of Experimental Psychology: Human Perception and Performance, 25*(3), 649.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19*(5), 448-458.

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review, 11*(4), 716-722.

Goldstein, A. G., Knight, P., Bailis, K., & Conover, J. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society, 17*(5), 217-220.

Goldstone, R. L. (1995). Effects of categorization on color perception. *Psychological Science, 6*(5), 298-304.

Goozée, J. V., Lapointe, L. L., & Murdoch, B. E. (2003). Effects of speaking rate on EMA-derived lingual kinematics: A preliminary investigation. *Clinical Linguistics & Phonetics, 17*(4-5), 375-381.

Gordon, M. S., Daneman, M., & Schneider, B. A. (2009). Comprehension of speeded discourse by younger and older listeners. *Experimental Aging Research, 35*(3), 277-296.

Graddol, D., & Swann, J. (1983). Speaking fundamental frequency: Some physical and social correlates. *Language and Speech, 26*(4), 351-366.

Green, D., & Swets, J. (1988). *Signal detection theory and psychophysics*. John Wiley & Sons, Inc, New York.

Green, J. R., & Nip, I. S. (2010). Some organization principles in early speech development. *Speech Motor Control: New Developments in Basic and Applied Research,* 171-188.

Greenberg, M. S., Westcott, D. R., & Bailey, S. E. (1998). When believing is seeing: The effect of scripts on eyewitness memory. *Law and Human Behavior, 22*(6), 685-694.

Guy, G., Horvath, B., Vonwiller, J., Daisley, E., & Rogers, I. (1986). An intonational change in progress in Australian English. *Language in Society, 15*(1), 23-51.

Halberstadt, J. B., & Niedenthal, P. M. (2001). Effects of emotion concepts on perceptual memory for emotional expressions. *Journal of Personality and Social Psychology, 81*(4), 587.

Hammersley, R., & Read, J. D. (1985). The effect of participation in a conversation on recognition and identification of the speakers' voices. *Law and Human Behavior, 9*(1), 71.

Hammersley, R., & Read, J. D. (1996). Voice identification by humans and computers. In S. L. Sporer, R. S. Malpass, & G. Koehnken (Eds.), *Psychological issues in eyewitness identification* (pp. 117-152). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hanson, V. L. (1977). Within-category discriminations in speech perception. *Attention, Perception, & Psychophysics, 21*(5), 423-430.

Harnsberger, J. D., Shrivastav, R., Brown, W., Rothman, H., & Hollien, H. (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice, 22*(1), 58-69.

Harnsberger, J. D., Shrivastav, R., Brown, W., Rothman, H., & Hollien, H. (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice, 22*(1), 58-69.

Harries, M., Hawkins, S., Hacking, J., & Hughes, I. (1998). Changes in the male voice at puberty: Vocal fold length and its relationship to the fundamental frequency of the voice. *The Journal of Laryngology & Otology, 112*(5), 451-454.

Harrigan, J. A., Gramata, J. F., Lucic, K. S., & Margolis, C. (1989). It's how you say it: Physicians' vocal behavior. *Social Science & Medicine, 28*(1), 87-92.

Harrington, J., & Cassidy, S. (2012). *Techniques in speech acoustics.* Springer Science & Business Media, The Netherlands.

Harris, J. D. (1952). The decline of pitch discrimination with time. *Journal of Experimental Psychology, 43*(2), 96-99.

Hartman, D. E., & Danhauer, J. L. (1976). Perceptual features of speech for males in four perceived age decades. *The Journal of the Acoustical Society of America, 59*(3), 713-715.

Haselager, G., Slis, I., & Rietveld, A. (1991). An alternative method of studying the development of speech rate. *Clinical Linguistics & Phonetics*, *5*(1), 53-63.

Hashtroudi, S., Johnson, M. K., & Chrosniak, L. D. (1989). Aging and source monitoring. *Psychology and Aging, 4*(1), 106-112.

Haslam, S. A., & Turner, J. C. (1992). Context-dependent variation in social stereotyping 2: The relationship between frame of reference, self-categorization and accentuation. *European Journal of Social Psychology, 22*(3), 251-277.

Heffernan, K. (2007). The role of phonemic contrast in the formation of Sino-Japanese. *Journal of East Asian Linguistics, 16*(2), 61-86.

Heffernan, K. (2010). Mumbling is macho: Phonetic distinctiveness in the speech of American radio DJs. *American Speech, 85*(1), 67-90.

Henton, C. (1995). Pitch dynamism in female and male speech. *Language & Communication, 15*(1), 43-61.

Herlitz, A., Nilsson, L., & Bäckman, L. (1997). Gender differences in episodic memory. *Memory & Cognition, 25*(6), 801-811.

Hertrich, I., & Ackermann, H. (2000). Lip–jaw and tongue–jaw coordination during rate-controlled syllable repetitions. *The Journal of the Acoustical Society of America, 107*(4), 2236-2247.

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f 0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics, 71*(5), 1150-1166.

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f 0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics, 71*(5), 1150-1166.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099-3111.

Hilliar, K. F., & Kemp, R. I. (2008). Barack Obama or Barry Dunham? The appearance of multiracial faces is affected by the names assigned to them. *Perception, 37*(10), 1605-1608.

Hirano, M. (1974). Morphological structure of the vocal cord as a vibrator and its variations. *Folia Phoniatrica Et Logopaedica, 26*(2), 89-94.

Hirano, M., Kurita, S., & Nakashima, T. (1983). Growth, development and aging of human vocal folds. *Vocal Fold Physiology: Contemporary Research and Clinical Issues,* 22-43.

Hirano, M., Kurita, S., & Sakaguchi, S. (1989). Ageing of the vibratory tissue of human vocal folds. *Acta Oto-Laryngologica, 107*(5-6), 428-433.

Hixon, T. J., & Hoit, J. D. (2006). A clinical method for the detection and quantification of quick respiratory hyperkinesia. *American Journal of Speech-Language Pathology, 15*(1), 15-19.

Hixon, T. J., Weismer, G., & Hoit, J. D. (2008). *Preclinical speech science.* Plural Publishing Inc, England.

Hogg, M., & Vaughan, G. (2010). *Essentials of Social Psychology,* Pearson Education Limited, England.

Hollien, H., Saletto, J., & Miller, S. (1993). Psychological stress in voice: A new approach. *Studia Phonetica Posnaniensia, 4*, 5-17.

Hollien, H. (1990). *The acoustics of crime. The new science of forensic phonetics*. Springer US, America.

Hollien, H. (1996). Consideration of guidelines for earwitness lineups. *Forensic Linguistics, 3*, 14-23.

Hollien, H., Bennett, G., & Gelfer, M. (1983). Criminal identification comparison: Aural versus visual identifications resulting from a simulated crime. *Journal of Forensic Science, 28*(1), 208-221.

Hollien, H., Huntley, R., Kunzel, H., & Hollien, P. A. (2013). Criteria for earwitness lineups. *International Journal of Speech Language and the Law, 2*(2), 143-153.

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics, 10,* 139-148.

Hollien, H., & Shipp, T. (1972). Speaking fundamental frequency and chronologic age in males. *Journal of Speech, Language, and Hearing Research, 15*(1), 155-159.

Honorof, D. N., & Whalen, D. (2010). Identification of speaker sex from one vowel across a range of fundamental frequencies. *The Journal of the Acoustical Society of America, 128*(5), 3095-3104.

Hopp, S. L., Owren, M. J., & Evans, C. S. (Eds.). (2012). *Animal acoustic communication: sound analysis and research methods*. Springer Science & Business Media, United Kingdom.

Horii, Y., & Ryan, W. J. (1981). Fundamental frequency characteristics and perceived age of adult male speakers. *Folia Phoniatrica Et Logopaedica, 33*(4), 227-233.

Huart, J., Corneille, O., & Becquart, E. (2005). Face-based categorization, context-based categorization, and distortions in the recollection of gender ambiguous faces. *Journal of Experimental Social Psychology, 41*(6), 598-608.

Huber, J. E., & Spruill, J. (2008). Age-related changes to speech breathing with increased vocal loudness. *Journal of Speech, Language, and Hearing Research, 51*(3), 651-668.

Hughes, S. M., & Rhodes, B. C. (2010). Making age assessments based on voice: The impact of the reproductive viability of the speaker. *Journal of Social, Evolutionary, and Cultural Psychology, 4*(4), 290.

Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., . . . Hatano, K. (1997). Vocal identification of speaker and emotion activates differerent brain regions. *Neuroreport, 8*(12), 2809-2812.

Ingemann, F. (1968). Identification of the speaker's sex from voiceless fricatives. *The Journal of the Acoustical Society of America, 44*(4), 1142-1144.

Israel, H. (1973). Age factor and the pattern of change in craniofacial structures. *American Journal of Physical Anthropology, 39*(1), 111-128.

Ivry, R. B., & Spencer, R. M. (2004). The neural representation of time. *Current Opinion in Neurobiology*, *14*(2), 225-232.

Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America, 128*(2), 839-850.

Jansen, A., & Niyogi, P. (2006, May). Intrinsic Fourier analysis on the manifold of speech sounds. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* (Vol. 1, pp. I-I). IEEE.

Johnson, M. K., De Leonardis, D. M., Hashtroudi, S., & Ferguson, S. A. (1995). Aging and single versus multiple cues in source monitoring. *Psychology and Aging, 10*(4), 507.

Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., & Vukovic, J. (2010). A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour*, *79*(1), 57-62.

Jreige, C., Patel, R., & Bunnell, H. T. (2009). VocaliD: Personalizing text-to-speech synthesis for individuals with severe speech impairment. *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility,* 259-260.

Junger, J., Pauly, K., Bröhr, S., Birkholz, P., Neuschaefer-Rube, C., Kohler, C., . . . Habel, U. (2013). Sex matters: Neural correlates of voice gender perception. *Neuroimage, 79*, 275-287.

Kadakia, S. (2013). Care of the professional voice: The effect of hormones on the voice. *Journal of Singing-the Official Journal of the National Association of Teachers of Singing, 69*(5), 571-574.

Kahane, J. C. (1981). Anatomic and physiologic changes in the aging peripheral speech mechanism. *Aging: Communication Processes and Disorders,* 21-45.

Kent, R. D., & Moll, K. L. (1972). Cinefluorographic analyses of selected lingual consonants. *Journal of Speech, Language, and Hearing Research, 15*(3), 453-473.

Kent, R. D. (1984). Psychobiology of speech development: Coemergence of language and a movement system. *The American Journal of Physiology, 246*(6 Pt 2), R888-94.

Kerstholt, J. H., Jansen, N. J., Van Amelsvoort, A. G., & Broeders, A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology, 18*(3), 327-336.

Kerstholt, J. H., Jansen, N. J., Van Amelsvoort, A. G., & Broeders, A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology, 20*(2), 187-197.

Kinchla, R. (1973). Selective processes in sensory memory: A probe-comparison procedure. In S. Kornblum (Ed.), *Attention and Performance IV.* Academic Press, New York.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87*(2), 820-857.

Kleider, H. M., Pezdek, K., Goldinger, S. D., & Kirk, A. (2008). Schema-driven source misattribution errors: Remembering the expected from a witnessed event. *Applied Cognitive Psychology, 22*(1), 1-20.

Köster, O., & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics, 4*(1), 1350-1771.

Köster, O., Schiller, N. O., & Künzel, H. (1995). The influence of native-language background on speaker recognition. *Proceedings of the XIIIth International Congress of Phonetic Sciences, 4*, 306-309.

Kowal, S., O'Connell, D. C., & Sabin, E. J. (1975). Development of temporal patterning and vocal hesitations in spontaneous narratives. *Journal of Psycholinguistic Research, 4*(3), 195-207.

Kowalska, D. M. (1997). The method of training dogs in auditory recognition memory tasks with trial-unique stimuli. *Acta Neurobiologiae Experimentalis, 57*, 345-352.

Kramer, C. (1977). Perceptions of female and male speech. *Language and Speech, 20*(2), 151-161.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Wiley-Blackwell, United Kingdom.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology, 67*(4), 596.

Krueger, J., & Rothbart, M. (1990). Contrast and accentuation effects in category learning. *Journal of Personality and Social Psychology, 59*(4), 651.

Krumhansl, C. L. (2001). *Cognitive foundations of musical pitch.* Oxford University Press, United Kingdom.

Kuehn, D. P., & Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics, 4*(4), 303-320.

Kuwabara, H., & Takagi, T. (1991). Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method. *Speech Communication, 10*(5-6), 491-495.

Lachs, L., & Pisoni, D. B. (2004). Crossmodal source identification in speech perception. *Ecological Psychology, 16*(3), 159-187.

Ladefoged, P. (1962). Subglottal activity during speech. In *Proceedings of the Fourth International Congress of Phonetic Sciences*. The Hague, The Netherlands.

Ladefoged, P., & Ladefoged, J. (1980). The ability of listeners to identify voices. *UCLA Working Papers in Phonetics, 49,* 43-51.

Lampinen, J. M., Faries, J. M., Neuschatz, J. S., & Toglia, M. P. (2000). Recollections of things schematic: The influence of scripts on recollective experience. *Applied Cognitive Psychology, 14*(6), 543-554.

Lapointe, F. (2005). Choreogenetics: The generation of choreographic variants through genetic mutations and selection. *Proceedings of the 7th Annual Workshop on Genetic and Evolutionary Computation,* 366-369.

LaRiviere, C. L. (1972). Some Acoustic and Perceptual Correlates of Speaker Identification., *Dissertation Abstracts International, 32,* 12-12A.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *The Journal of the Acoustical Society of America, 59*(3), 675-678.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *The Journal of the Acoustical Society of America, 59*(3), 675-678.

Lass, N. J., Mertz, P. J., & Kimmel, K. L. (1978). The effect of temporal speech alterations on speaker race and sex identifications. *Language and Speech, 21*(3), 279-290.

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology, 21*(4), R143-R145.

Latinus, M., & Belin, P. (2012). Perceptual auditory aftereffects on voice identity using brief vowel stimuli. *PLoS One, 7*(7), e41384.

Laver, J. (1979). The description of voice quality in general phonetic theory. *Work Prog-University Edinburgh, Department of Linguistics, 12*, 30-52.

Laver, J. (1989). Cognitive science and speech: A framework for research. *Logic and Linguistics: Research Directions in Cognitive Science European Perspective. Hillsdale, NJ: Lawrence Erlbaum,* 37-70.

Laver, J. (1991). *The gift of speech: Papers in the analysis of speech and voice*. Edinburgh University Press, United Kingdom.

Laver, J. (1994). Principles of phonetics. *Cambridge University Press, Cambridge, United Kingdom.*

Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K. Scherer & H. Giles (Eds.), Social Markers in Speech (pp. 1-32). Cambridge University Press, Cambridge, UK.

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication, 30*(1), 9-26.

Legge, G. E., Grosmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(2), 298.

Lehiste, I. (1976). Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics, 4*(2), 113-117.

Levi, S. V., & Pisoni, D. B. (2007). Indexical and linguistic channels in speech perception: Some effects of voiceovers on advertising outcomes. *Psycholinguistic Phenomena in Marketing Communications,* 203-219.

Levi, S. V., Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible? *The Journal of the Acoustical Society of America, 130*(6), 4053-4062.

Levin, D. T. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General, 129*(4), 559.

Levin, D. T., & Banaji, M. R. (2006). Distortions in the perceived lightness of faces: The role of race categories. *Journal of Experimental Psychology: General, 135*(4), 501.

Lewin, C., Wolgers, G., & Herlitz, A. (2001). Sex differences favoring women in verbal but not in visuospatial episodic memory. *Neuropsychology, 15*(2), 165.

Lieberman, P., Laitman, J. T., Reidenberg, J. S., & Gannon, P. J. (1992). The anatomy, physiology, acoustics and perception of speech: Essential elements in analysis of the evolution of human speech. *Journal of Human Evolution, 23*(6), 447-467.

Linville, S. E. (1992). Glottal gap configurations in two age groups of women. *Journal of Speech, Language, and Hearing Research, 35*(6), 1209-1215.

Linville, S. E. (1996). The sound of senescence. *Journal of Voice, 10*(2), 190-200.

Linville, S. E., & Fisher, H. B. (1985). Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. *The Journal of the Acoustical Society of America, 78*(1), 40-48.

Linville, S. E., & Rens, J. (2001). Vocal tract resonance analysis of aging voice using long-term average spectra. *Journal of Voice, 15*(3), 323-330.

Livingstone, S. R., Thompson, W. F., Wanderley, M. M., & Palmer, C. (2015). Common cues to emotion in the dynamic facial expressions of speech and song. *The Quarterly Journal of Experimental Psychology, 68*(5), 952-970.

Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about" weapon focus.". *Law and Human Behavior, 11*(1), 55.

Loveday, L. (1981). Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of english and japanese politeness formulae. *Language and Speech, 24*(1), 71-89.

Lu, Z., Williamson, S., & Kaufman, L. (1992). Behavioral lifetime of human auditory sensory memory predicted by physiological measures. *Science,* 1668-1670.

MacLin, O. H., & Malpass, R. S. (2001). Racial categorization of faces: The ambiguous race face effect. *Psychology, Public Policy, and Law, 7*(1), 98.

Marcell, M. M., Borella, D., Greene, M., Kerr, E., & Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology, 22*(6), 830-864.

Margolis, H. (1987). *Patterns, thinking, and cognition: A theory of judgment* University of Chicago Press, United States of America.

Martin, J. R., & Rose, D. (2003). *Working with discourse: Meaning beyond the clause* Bloomsbury Publishing.

Masthoff, H. (1996). A report on a voice disguise experiment. *Forensic Linguistics, 3,* 160-167.

Mathur, S., Choudhary, S., & Vyas, J. (2013). Speaker recognition system and its forensic implications. *Open Access Scientific Reports, 2*(4), 1-6.

Mauk, M. D., & Buonomano, D. V. (2004). The neural basis of temporal processing. *Annual Review of Neuroscience., 27*, 307-340.

Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance, 39*(2), 307.

McGivern, R. F., Huston, J. P., Byrd, D., King, T., Siegle, G. J., & Reilly, J. (1997). Sex differences in visual recognition memory: Support for a sex-related difference in attention in adults and children. *Brain and Cognition, 34*(3), 323-336.

McGlone, R. E., & Hollien, H. (1963). Vocal pitch characteristics of aged women. *Journal of Speech, Language, and Hearing Research*, *6*(2), 164-170.

McGowan, R. S. (1994). Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication, 14*(1), 19-48.

McKinney, J. C. (1994). *The diagnosis and correction of vocal faults: A manual for teachers of singing and for choir directors.* Waveland Press, Inc, United States of America.

Meissner, C. A., & Brigham, J. C. (2001). Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review. *Psychology, Public Policy, and Law, 7*(1), 33-35.

Michelsson, K., Eklund, K., Leppanen, P., & Lyytinen, H. (2002). Cry characteristics of 172 healthy 1-to 7-day-old infants. *Folia Phoniatrica Et Logopaedica: Official Organ of the International Association of Logopedics and Phoniatrics (IALP), 54*(4), 190-200.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica, 41*(4), 215-225.

Mondor, T. A., Hurlburt, J., & Thorne, L. (2003). Categorizing sounds by pitch: Effects of stimulus similarity and response repetition. *Attention, Perception, & Psychophysics, 65*(1), 107-114.

Moore, B. C. J. (1995). *Hearing.* Academic Press, San Diego.

Mount, K. H., & Salmon, S. J. (1988). Changing the vocal characteristics of a postoperative transsexual patient: A longitudinal study. *Journal of Communication Disorders, 21*(3), 229-238.

Muchnik, C., Efrati, M., Nemeth, E., Malin, M., & Hildesheimer, M. (1991). Central auditory skills in blind and sighted subjects. *Scandinavian Audiology, 20*(1), 19-23.

Mullennix, J. W., Stern, S. E., Grounds, B., Kalas, R., Flaherty, M., Kowalok, S., . . . Tessmer, B. (2010). Earwitness memory: Distortions for voice pitch and speaking rate. *Applied Cognitive Psychology, 24*(4), 513-526.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America, 93*(2), 1097-1108.

Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *The Journal of the Acoustical Society of America, 68*(5), 1294-1300.

Neuschatz, J. S., Lampinen, J. M., Preston, E. L., Hawkins, E. R., & Toglia, M. P. (2002). The effect of memory schemata on memory and the phenomenological experience of naturalistic situations. *Applied Cognitive Psychology, 16*(6), 687-708.

Newman, S., Butler, J., Hammond, E. H., & Gray, S. D. (2000). Preliminary report on hormone receptors in the human vocal fold. *Journal of Voice, 14*(1), 72-81.

Nip, I. S., & Green, J. R. (2013). Increases in cognitive and linguistic processing primarily account for increases in speaking rate with age. *Child Development, 84*(4), 1324-1337.

Nolan, F. (2005). Forensic speaker identification and the phonetic description of voice quality. In W. J. Hardcastle & J. Mackenzie Beck (Ed.), *A Figure of Speech* (pp. 385-411). Routledge, New York.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Attention, Perception, & Psychophysics, 60*(3), 355-376.

Oates, J., & Dacakis, G. (1997). Voice change in transsexuals. *Venereology, 10*(3), 178.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica, 41*(1), 1-16.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica, 41*(1), 1-16.

Ohde, R. N., Sharf, D. J., & Jacobson, P. F. (1992). Phonetic analysis of normal and abnormal speech. *The Journal of the Acoustical Society of America, 92*(6), 3452-3452.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434-447.

Orchard, T. L., & Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology, 9*(3), 249-260.

Ormerod, D. (2001). Sounds familiar? voice identification evidence. *Criminal Law Review,* 595-622.

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *The Journal of the Acoustical Society of America, 77*(2), 640-648.

Overbeck, J. L. (2005). Beyond admissibility: A practical look at the use of eyewitness expert testimony in the federal courts. *NYUL Rev., 80*, 1895.

Owren, M. J., Berkowitz, M., & Bachorowski, J. (2007). Listeners judge talker sex more efficiently from male than from female vowels. *Attention, Perception, & Psychophysics, 69*(6), 930-941.

Palmer, S. D., Havelka, J., & van Hooff, J. C. (2013). Changes in recognition memory over time: An ERP investigation into vocabulary learning. *PloS One, 8*(9), e72870.

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician enhancement for speech-in-noise. *Ear and hearing, 30*(6), 653-661.

Peirce, J. W. (2007). PsychoPy—psychophysics software in python. *Journal of Neuroscience Methods, 162*(1), 8-13.

Pépiot, E. (2014). Male and female speech: A study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers. *Speech Prosody 7,* 305-309.

Pépiot, E. (2015). Voice, speech and gender: male-female acoustic differences and cross-language variation in English and French speakers. *Corela. Cognition, Représentation, Langage,* (HS-16).

Perry, T. L., Ohde, R. N., & Ashmead, D. H. (2001). The acoustic bases for gender identification from children’s voices. *The Journal of the Acoustical Society of America, 109*(6), 2988-2998.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*(2), 175-184.

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology, 21*(4), 539-550.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Attention, Perception, & Psychophysics, 13*(2), 253-260.

Plack, C. J. (2013). *The sense of hearing*. Psychology Press, New York.

Ptacek, P. H., & Sander, E. K. (1966). Age recognition from voice. *Journal of Speech and Hearing Research, 9*(2), 273-277.

Puts, D. A. (2005). Mating context and menstrual phase affect women's preferences for male voice pitch. *Evolution and Human Behavior, 26*(5), 388-397.

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior, 27*(4), 283-296.

Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. (2007). Men's voices as dominance signals: Vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior, 28*(5), 340-344.

Puts, D. A., Apicella, C. L., & Cardenas, R. A. (2012). Masculine voices signal men's threat potential in forager and industrial societies. *Proceedings.Biological Sciences, 279*(1728), 601-609.

Queller, S., Schell, T., & Mason, W. (2006). A novel view of between-categories contrast and within-category assimilation. *Journal of Personality and Social Psychology, 91*(3), 406.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America, 123*(2), 1104-1113.

Radvansky, G. A. (2015). *Human memory.* Psychology Press, New York.

Ramig, L. A. (1986). Aging speech: Physiological and sociological aspects. *Language & Communication, 6*(1-2), 25-34.

Rastatter, M. P., & Jacques, R. D. (1990). Formant frequency structure of the aging male and female vocal tract. *Folia Phoniatrica Et Logopaedica, 42*(6), 312-319.

Read, D., & Craik, F. I. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied, 1*(1), 6-18.

Reby, D., & McComb, K. (2003). Anatomical constraints generate honesty: Acoustic cues to age and weight in the roars of red deer stags. *Animal Behaviour, 65*(3), 519-530.

Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the Acoustical Society of America, 66*(4), 1023-1028.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance, 23*(3), 651.

Rendall, D., Vokey, J. R., & Nemeth, C. (2007). Lifting the curtain on the wizard of oz: Biased voice-based impressions of speaker size. *Journal of Experimental Psychology: Human Perception and Performance, 33*(5), 1208-1219.

Riely, R. R., & Smith, A. (2003). Speech movements do not scale by orofacial structure size. *Journal of Applied Physiology (Bethesda, Md.: 1985), 94*(6), 2119-2126.

Robb, M. P., Maclagan, M. A., & Chen, Y. (2004). Speaking rates of American and New Zealand varieties of English. *Clinical Linguistics & Phonetics, 18*(1), 1-15.

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PloS one, 11*(2), e0150036.

Rodman, R. (1998). Speaker recognition of disguised voices: A program for research. In *Proceedings of the 8th COST 250 workshop, Ankara: Speaker identification by man and by machine: Directions for forensic applications* (pp. 9-22).

Roebuck, R., & Wilding, J. (1993). Effects of vowel variety and sample length on identification of a speaker in a line-up. *Applied Cognitive Psychology, 7*(6), 475-481.

Rose, P., & Duncan, S. (2013). Naïve auditory identification and discrimination of similar voices by familiar listeners. *International Journal of Speech Language and the Law, 2*(1), 1-17.

Rother, P., Wohlgemuth, B., Wolff, W., & Rebentrost, I. (2002). Morphometrically observable aging changes in the human tongue. *Annals of Anatomy-Anatomischer Anzeiger, 184*(2), 159-164.

Ryan, W., & Burk, K. (1974). Perceptual and acoustic correlates of aging in the speech of males. *Journal of Communication Disorders, 7*(2), 181-192.

Sachs, J., Lieberman, P., & Erickson, D. (1973). Anatomical and cultural determinants of male and female speech. In R. W. Shuy & R. W. Fasold (Eds.), *Language attitudes: Current trends and prospects* (pp. 74-84) Georgetown University Press, Washington.

Samuelsson, Y. (2006). Gender effects on phonetic variation and speaking styles. A Literature Study. *GSLT Speech Technology Term Paper.* Retrieved from http://www.speech.kth.se/~rolf/NGSLT/gslt_papers_2006/YvonneStermpaper.pdf.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology, 65*(1), 111.

Sataloff, R. T. (2006). *Vocal health and pedagogy, volume II: Advanced assessment and practice.* Plural Publishing, North America.

Sbattella, L., Colombo, L., Rinaldi, C., Tedesco, R., Matteucci, M., & Trivilini, A. (2014). Extracting emotions and communication styles from vocal signals. *PhyCS,* 183-195.

Scheibel, M. E., Tomiyasu, U., & Scheibel, A. B. (1977). The aging human betz cell. *Experimental Neurology, 56*(3), 598-609.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1), 227-256.

Scherer, K. R., London, H., & Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality, 7*(1), 31-44.

Schiller, N. O., & Köster, O. (1996). Evaluation of a foreign speaker in forensic phonetics: A report. *Forensic Linguistics: The International Journal of Speech, Language and the Law, 3*, 176-185.

Schmidt-Nielsen, A., & Stern, K. R. (1985). Identification of known voices as a function of familiarity and narrow-band coding. *The Journal of the Acoustical Society of America, 77*(2), 658-663.

Schneiderman, H., & Kanade, T. (1998). Probabilistic modelling of local appearance and spatial relationships for object recognition. *Computer Vision and Pattern Recognition*.45-55.

Schwartz, M. F., & Rine, H. E. (1968). Identification of speaker sex from isolated, whispered vowels. *The Journal of the Acoustical Society of America, 44*(6), 1736-1737.

Sell, G., Suied, C., Elhilali, M., & Shamma, S. (2015). Perceptual susceptibility to acoustic manipulations in speaker discrimination. *The Journal of the Acoustical Society of America, 137*(2), 911-922.

Sell, G., Suied, C., Elhilali, M., & Shamma, S. (2015). Perceptual susceptibility to acoustic manipulations in speaker discrimination. *The Journal of the Acoustical Society of America, 137*(2), 911-922.

Shaiman, S. (2001). Kinematics of compensatory vowel shortening: The effect of speaking rate and coda composition on intra-and inter-articulatory timing. *Journal of Phonetics, 29*(1), 89-107.

Shaiman, S. (2002). Articulatory control of vowel length for contiguous jaw cycles: The effects of speaking rate and phonetic context. *Journal of Speech, Language, and Hearing Research, 45*(4), 663.

Shaiman, S., Adams, S. G., & Kimelman, M. D. (1995). Timing relationships of the upper lip and jawacross changes in speaking rate. *Journal of Phonetics*, *23*(1), 119-128.

Sheffert, S. M., & Olson, E. (2004). Audiovisual speech facilitates voice learning. *Attention, Perception, & Psychophysics*, *66*(2), 352-362.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance, 28*(6), 1447.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance, 28*(6), 1447.

Sherrin, C. (2014). Earwitness evidence: The reliability of voice identifications. *Osgoode Legal Studies Research Paper Series, 52,* 2-44.

Shipp, T., Qi, Y., Huntley, R., & Hollien, H. (1992). Acoustic and temporal correlates of perceived age. *Journal of Voice, 6*(3), 211-216.

Shrivastav, R., Hollien, H., Brown Jr, W., Rothman, H. B., & Harnsberger, J. D. (2003). Shifting perceptions of age in voice. *The Journal of the Acoustical Society of America, 114*(4), 2336-2337.

Siegman, A. W., & Boyle, S. (1993). Voices of fear and anxiety and sadness and depression: The effects of speech rate and loudness on fear and anxiety and sadness and depression. *Journal of Abnormal Psychology, 102*(3), 430.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass, 3*(2), 621-640.

Simpson, A. P., & Ericsdotter, C. (2003). Sex-specific durational differences in English and Swedish. *Proc. XVth ICPhS,* 1113-1116.

Simpson, A., & Ericsdotter, C. (2007). Sex-specific differences in f0 and vowel space. *XVIth International Congress of Phonetic Sciences,* 933-936.

Singh, S., & Murry, T. (1978). Multidimensional classification of normal voice qualities. *The Journal of the Acoustical Society of America, 64*(1), 81-87.

Skoog Waller, S., Eriksson, M., & Sorqvist, P. (2015). Can you hear my age? influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology, 6*, 978-989.

Skuk, V. G., & Schweinberger, S. R. (2013). Adaptation aftereffects in vocal emotion perception elicited by expressive faces and voices. *PloS One, 8*(11), e81691.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? *Psychological Science*, *17*(8), 675-681.

Smith, A., & Goffman, L. (1998). Stability and patterning of speech movement sequences in children and adults. *Journal of Speech, Language, and Hearing Research, 41*(1), 18-30.

Smith, A., & Zelaznik, H. N. (2004). Development of functional synergies for speech motor coordination in childhood and adolescence. *Developmental Psychobiology, 45*(1), 22-33.

Smith, B. L., & Gartenberg, T. E. (1984). Initial observations concerning developmental characteristics of labio-mandibular kinematics. *The Journal of the Acoustical Society of America, 75*(5), 1599-1605.

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age a. *The Journal of the Acoustical Society of America, 118*(5), 3177-3186.

Smith, D. R., Walters, T. C., & Patterson, R. D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled a. *The Journal of the Acoustical Society of America, 122*(6), 3628-3639.

Smith, H. M., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016). The effect of inserting an inter-stimulus interval in face–voice matching tasks. *The Quarterly Journal of Experimental Psychology,* 1-12.

Sokhi, D. S., Hunter, M. D., Wilkinson, I. D., & Woodruff, P. W. (2005). Male and female voices activate distinct regions in the male brain. *Neuroimage, 27*(3), 572-578.

Solan, L., & Tiersma, P. (2003). Falling on deaf ears: Scientists say that earwitnesses are unreliable. why aren't courts listening. *Legal Affairs, 71*(2), 7-19.

Spencer, L. E. (1988). Speech characteristics of male-to-female transsexuals: A perceptual and acoustic study. *Folia Phoniatrica Et Logopaedica, 40*(1), 31-42.

Stern, S. E., Mullennix, J. W., Corneille, O., & Huart, J. (2007). Distortions in the memory of the pitch of speech. *Experimental Psychology, 54*(2), 148-160.

Stevens, K. N. (1980). Acoustic correlates of some phonetic categories. *The Journal of the Acoustical Society of America, 68*(3), 836-842.

Stewart, M. A., & Ryan, E. B. (1982). Attitudes toward younger and older adult speakers: Effects of varying speech rates. *Journal of Language and Social Psychology, 1*(2), 91-109.

Story, B. H., & Titze, I. R. (1995). Voice simulation with a body-cover model of the vocal folds. *The Journal of the Acoustical Society of America, 97*(2), 1249-1260.

Strauss, R. H., Liggett, M. T., & Lanese, R. R. (1985). Anabolic steroid use and perceived effects in ten weight-trained women athletes. *Jama, 253*(19), 2871-2873.

Street Jr, R. L., Brady, R. M., & Putman, W. B. (1983). The influence of speech rate stereotypes and rate similarity or listeners' evaluations of speakers. *Journal of Language and Social Psychology, 2*(1), 37-56.

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British Journal of Psychology, 54*(2), 101-114.

Tanaka, J. W. (2001). The entry point of face recognition: Evidence for face expertise. *Journal of Experimental Psychology-General*, *130*(3), 534-543.

Taylor, A. M., Reby, D., & McComb, K. (2011). Cross modal perception of body size in domestic dogs (canis familiaris). *PLoS One, 6*(2), e17069.

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology, 1*(2), 121-131.

Thurman, L., & Welch, G. (2000). *Bodymind & voice: Foundations of voice education*. The VoiceCare Network, United States of America.

Titze, I. R. (1994). *Principles of voice production.* Allyn & Bacon Publishing, United States.

Titze, I. R. (1989). On the relation between subglottal pressure and fundamental frequency in phonation. *The Journal of the Acoustical Society of America*, *85*(2), 901-906.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America, 85*(4), 1699-1707.

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America, 107*(6), 3438-3451.

Treisman, A. (1964). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior, 3*(6), 449-459.

Tsao, Y., & Weismer, G. (1997). Interspeaker variation in habitual speaking rate evidence for a neuromuscular component. *Journal of Speech, Language, and Hearing Research, 40*(4), 858-866.

Van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and dutch women. *Language and Speech, 38*(3), 253-265.

Van Lancker, D., Kreiman, J., & Wickens, T. (1985). Familiar voice recognition: Parameters and patterns. part II. recognition of rate-altered voices. *Journal of Phonetics, 13,* 39-52.

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia, 25*(5), 829-834.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and. *Journal of Phonetics, 13,* 39-52.

Van Wallendael, L. R., Surace, A., Parsons, D. H., & Brown, M. (1994). 'Earwitness' voice recognition: Factors affecting accuracy and impact on jurors. *Applied Cognitive Psychology, 8*(7), 661-677.

Van Wassenhove, V. (2009). Minding time in an amodal representational space. *Philosophical Transactions of the Royal Society of London.Series B, Biological Sciences, 364*(1525), 1815-1830.

Vanags, T., Carroll, M., & Perfect, T. J. (2005). Verbal overshadowing: A sound theory in voice recognition? *Applied Cognitive Psychology, 19*(9), 1127-1144.

Vernet, M., Martin, F., Baudouin, J., Tiberghien, G., & Franck, N. (2007). Visual pattern recognition: What makes faces so special? In K. B. Leeland (Ed.), *Face Recognition: New Research.* Nova Publishers, New York.

Vorperian, H. K., Kent, R. D., Gentry, L. R., & Yandell, B. S. (1999). Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: Preliminary results. *International Journal of Pediatric Otorhinolaryngology, 49*(3), 197-206.

Walker, J. F., Archibald, L. M., Cherniak, S. R., & Fish, V. G. (1992). Articulation rate in 3- and 5-year-old children. *Journal of Speech, Language, and Hearing Research, 35*(1), 4-13.

Waller, S. S., & Eriksson, M. (2016). Vocal age disguise: The role of fundamental frequency and speech rate and its perceived effects. *Frontiers in Psychology, 7,* 1814.

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal adaptation in right posterior superior temporal sulcus during face-voice emotional integration. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 34*(20), 6813-6821.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(3), 387-401.

Weirich, M., & Simpson, A. P. (2014). Differences in acoustic vowel space and the perception of speech tempo. *Journal of Phonetics, 43,* 1-10.

Weismer, G., Liss, J., Moore, C., Yorkston, K., & Beukelman, D. (1991). Reductionism is a dead-end in speech research: Perspectives on a new direction. *Dysarthria and Apraxia of Speech: Perspectives on Management*, 15-28.

Wells, J. (1962, March). *A Study of the Formants of the Pure Vowels of British English.* Retrieved from http://www.phon.ucl.ac.uk/home/wells/formants/index.htm.

Wenndt, S. J. (2016). Human recognition of familiar voices. *The Journal of the Acoustical Society of America, 140*(2), 1172-1183.

Whiteside, S. P. (1996). Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences. *Journal of the International Phonetic Association, 26*(1), 23-40.

Whiteside, S. P. (1996). Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences. *Journal of the International Phonetic Association, 26*(1), 23-40.

Whiteside, S. P. (1998). The identification of a speaker's sex from synthesized vowels. *Perceptual and Motor Skills, 87*(2), 595-600.

Whiteside, S. P. (2001). Sex-specific fundamental and formant frequency patterns in a cross-sectional study. *The Journal of the Acoustical Society of America, 110*(1), 464-478.

Wickelgren, W. A. (1969). Associative strength theory of recognition memory for pitch. *Journal of Mathematical Psychology, 6*(1), 13-61.

Wilding, J., Cook, S., & Davis, J. (2000). Sound familiar. *The Psychologist, 13*(11), 558-562.

Willott, J. F. (1999). *Neurogerontology: Aging and the nervous system.* Springer Publishing Company, United States of America.

Winkler, R. (2007). Influences of pitch and speech rate on the perception of age from voice. *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken,* 1849-1852.

Winograd, E., Kerr, N. H., & Spence, M. J. (1984). Voice recognition: Effects of orienting task, and a test of blind versus sighted listeners. *The American Journal of Psychology,* 57-70.

Wolfe, V. I., Ratusnik, D. L., Smith, F. H., & Northrop, G. (1990). Intonation and fundamental frequency in male-to-female transsexuals. *J Speech Hear Disord, 55*(1), 43-50.

Xu, M., Homae, F., Hashimoto, R., & Hagiwara, H. (2013). Acoustic cues for the recognition of self-voice and other-voice. *Frontiers in Psychology, 4*, 735.

Xu, Y., Krishnan, A., & Gandour, J. T. (2006). Specificity of experience-dependent pitch representation in the brainstem. *Neuroreport, 17*(15), 1601-1605.

Yamazawa, H., & Hollien, H. (1992). Speaking fundamental frequency patterns of Japanese women. *Phonetica, 49*(2), 128-140.

Yarmey, A. D. (1986). Verbal, visual, and voice identification of a rape suspect under different levels of illumination. *Journal of Applied Psychology, 71*(3), 363.

Yarmey, A. D. (1991). Voice identification over the telephone. *Journal of Applied Social Psychology, 21*(22), 1868-1876.

Yarmey, A. D. (2000). Retrospective duration estimations for variant and invariant events in field situations. *Applied Cognitive Psychology, 14*(1), 45-57.

Yarmey, A. D. (2003). Earwitness identification over the telephone and in field settings. *Forensic Linguistics, 10*, 62-74.

Yarmey, A. D. (2004). Eyewitness recall and photo identification: A field experiment. *Psychology, Crime and Law, 10*(1), 53-68.

Yarmey, A. D., & Matthys, E. (1992). Voice identification of an abductor. *Applied Cognitive Psychology, 6*(5), 367-377.

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology, 15*(3), 283-299.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*(1), 23-43.

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences, 17*(6), 263-271.

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. *Proceedings of ICPhS XVI,* 1337-1340.

Zaske, R., Volberg, G., Kovacs, G., & Schweinberger, S. R. (2014). Electrophysiological correlates of voice learning and recognition. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 34*(33), 10821-10831.

APPENDIX A1. FUNDAMENTAL FREQUENCY (F0) VALUES FOR TARGET AND DISTRACTOR VOICES: EXPERIMENTS 5a, 6, AND 7 (Sentence One)

Table A1 illustrates the fundamental frequency (F0) values of all target and distractor voices for the spoken sentence *"Spring is the season where flowers appear, summer is the warmest season of the year"* used in Experiments 5a, 6, and 7.

Table A1: *Table displaying mean Fundamental Frequency (F0, measured in Hz) of all target and distractor voices for the spoken sentence "Spring is the season where flowers appear, summer is the warmest season of the year" used in Experiments 5a, 6, and 7.*

Voice	Target Voice (Hz)	Distractor	Fundamental Frequency (Hz)
Voice One (M)	High (116)	-10%	104
		-7%	108
		-5%	110
		+5%	122
		+7%	124
		+10%	128
	Moderate (106)	-10 %	95
		-7%	99
		-5%	101
		+5%	111
		+7%	113
		+10%	117
	Low (95)	-10%	86
		-7%	88
		-5%	90
		+5%	100
		+7 %	102

		+10%	105
Voice Two (M)	High (123)	-10%	111
		- 7%	114
		-5%	117
		+5%	129
		+7%	132
		+10%	135
	Moderate (112)	-10%	101
		-7%	104
		-5%	106
		+5%	118
		+7	120
		+10	123
	Low (101)	-10%	91
		-7%	94
		-5%	95
		+5%	106
		+7%	108
		+10%	111
Voice Three (F)	High (228)	-10%	205
		-7%	212
		-5%	217
		+5%	239
		+7%	244
		+10%	251
	Moderate (207)	-10%	186
		-7%	193
		-5%	197
		+5%	217
		+7%	221
		+10%	228

	Low (186)	-10%	167
		-7%	173
		-5%	177
		+5%	195
		+7%	199
		+10%	205
Voice Four (F)	High (238)	-10%	214
		-7%	221
		-5%	226
		+5%	250
		+7%	255
		+10%	262
	Moderate (217)	-10%	195
		-7%	202
		-5%	206
		+5%	228
		+7%	232
		+10%	239
	Low (195)	-10%	178
		-7%	181
		-5%	185
		+5%	205
		+7%	209
		+10%	215

Note: M = Male Voice; F = Female Voice. Fundamental Frequency (F0), measured in Hertz (Hz).

APPENDIX A2. SPEECH RATE VALUES FOR TARGET AND DISTRACTOR VOICES: EXPERIMENT 5b (Sentence One)

Table A2 illustrates the fundamental frequency (F0) values of all target and distractor voices for the spoken sentence *"Spring is the season where flowers appear, summer is the warmest season of the year"* used in Experiment 5b.

Table A2: *Table displaying the mean speech rate (syll/sec) of all target and distractor voices used in the present study for the spoken sentence "Spring is the season where flowers appear, summer is the warmest season of the year".*

Voice	Target Voice (syll/sec)	Distractor	Speech Rate (syll/sec)
Voice One (M)	Fast (4.31)	-20%	3.45
		-12%	3.79
		-10%	3.88
		+10%	4.74
		+12%	4.83
		+20%	5.17
	Moderate (3.59)	-20 %	2.87
		-12%	3.16
		-10%	3.23
		+10%	3.95
		+12%	4.02
		+20%	4.31
	Slow (2.87)	-20%	2.30
		-12%	2.53
		-10%	2.58
		+10%	3.16
		+12 %	3.21
		+20%	3.44

Voice Two (M)	Fast (4.31)	-20%	3.45
		-12%	3.79
		-10%	3.88
		+10%	4.74
		+12%	4.83
		+20%	5.17
	Moderate (3.26)	-20%	2.61
		-12%	2.87
		-10%	2.93
		+10%	3.59
		+12%	3.65
		+20%	3.91
	Slow (2.60)	-20%	2.08
		-12%	2.34
		-10%	2.29
		+10%	2.86
		+12%	2.91
		+20%	3.12
Voice Three (F)	Fast (4.33)	-20%	3.46
		-12%	3.81
		-10%	3.90
		+10%	4.76
		+12%	4.85
		+20%	5.20
	Moderate (3.62)	-20%	2.90
		-12%	3.19
		-10%	3.26
		+10%	3.98
		+12%	4.05
		+20%	4.34
	Slow (2.88)	-20%	2.30
		-12%	2.53

		-10%	2.59
		+10%	3.17
		+12%	3.23
		+20%	3.46
Voice Four (F)	Fast (3.80)	-20%	3.04
		-12%	3.34
		-10%	3.42
		+10%	4.18
		+12%	4.26
		+20%	4.56
	Moderate (3.17)	-20%	2.53
		-12%	2.78
		-10%	2.85
		+10%	3.49
		+12%	3.55
		+20%	3.80
	Slow (2.54)	-20%	2.03
		-12%	2.24
		-10%	2.29
		+10%	2.79
		+12%	2.84
		+20%	3.05

Note: M = Male Voice; F = Female Voice. Speech Rate measure in syllables per second (syll/sec).

APPENDIX A3. FUNDAMENTAL FREQUENCY (F0) VALUES FOR TARGET AND DISTRACTOR VOICES: EXPERIMENT 7

(Sentence Two)

Table A3 illustrates the fundamental frequency (F0) values of all target and distractor voices for the spoken sentence *"Living costs have more than tripled, and gas has gone down one third"* used in Experiment 7.

Table A3: *Table displaying mean Fundamental Frequency (F0, measured in Hz) of all target and distractor voices for the spoken sentence "Living costs have more than tripled, and gas has gone down one third" used in Experiment 7.*

Voice	Target Voice (Hz)	Distractor	Fundamental Frequency (Hz)
Voice One (M)	High (119)	-10%	107
		-5%	113
		+5%	125
		+10%	131
	Moderate (108)	-10 %	97
		-5%	103
		+5%	113
		+10%	119
	Low (97)	-10%	87
		-5%	92
		+5%	102
		+10%	107
Voice Two (M)	High (125)	-10%	113
		-5%	119
		+5%	131
		+10%	138

	Moderate (114)	-10%	103
		-5%	108
		+5%	120
		+10	125
	Low (103)	-10%	93
		-5%	98
		+5%	108
		+10%	113
Voice Three (F)	High (227)	-10%	204
		-5%	216
		+5%	238
		+10%	250
	Moderate (206)	-10%	185
		-5%	196
		+5%	216
		+10%	227
	Low (185)	-10%	167
		-5%	176
		+5%	194
		+10%	204
Voice Four (F)	High (232)	-10%	209
		-5%	220
		+5%	244
		+10%	255
	Moderate (211)	-10%	190
		-5%	200
		+5%	222
		+10%	232
	Low (190)	-10%	171
		-5%	181
		+5%	200

	+10%	209

Note: M = Male Voice; F = Female Voice. Fundamental Frequency (F0), measured in Hertz (Hz).

APPENDIX B1. ADDITIONAL FIGURES FOR EXPERIMENT 1a: FUNDAMENTAL FREQUENCY (F0)

Figure B1 illustrates the mean percentage of times listeners heard the manipulated versions of the voices as sounding the same as the original (i.e., unmanipulated) versions before the data was collapsed across plus and minus manipulations, for F0.

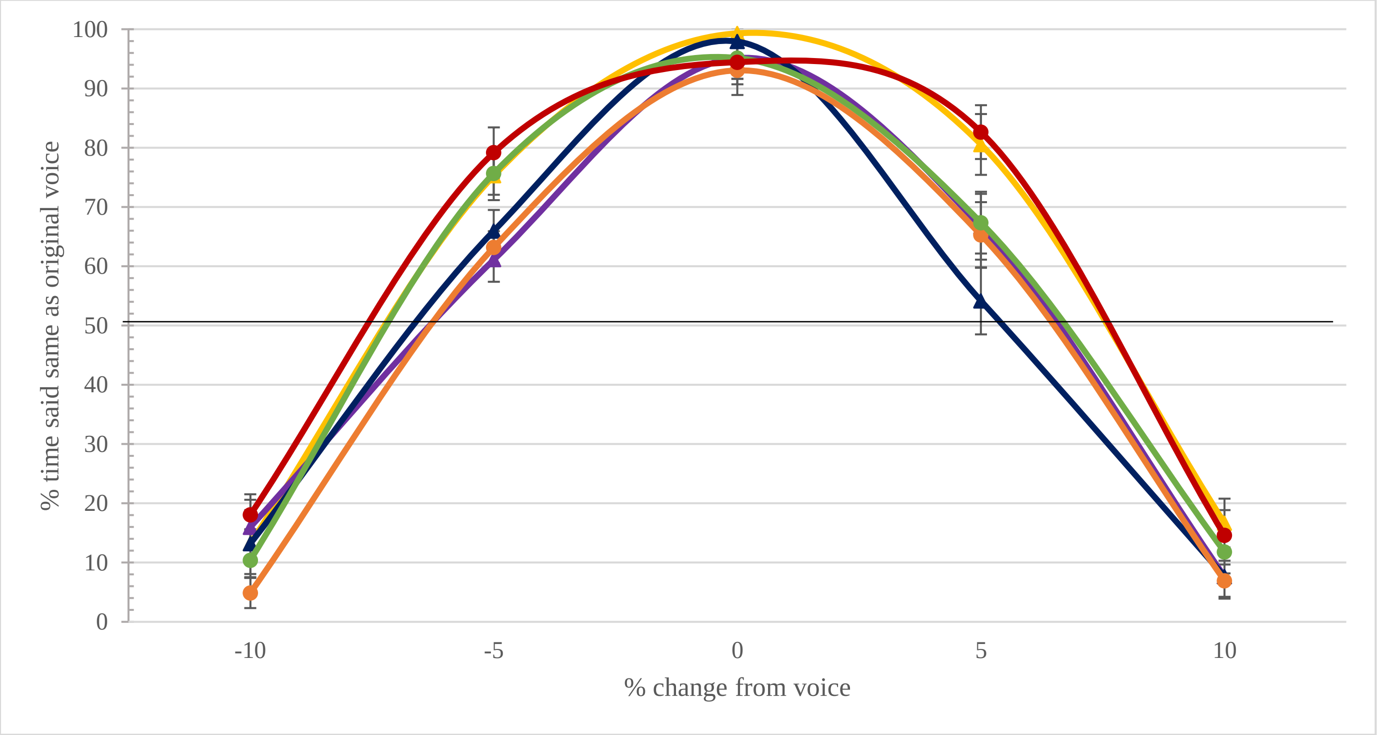


Figure B1: Line graph depicting percentage of times listeners heard voice as sounding the 'same' as the original voice, for F0. Each of the six voices are depicted by a different line colour. Male voices are depicted by the triangle symbol and female voices are depicted by the circle symbol. 95% confidence intervals are also shown.

APPENDIX B2. ADDITIONAL FIGURES FOR EXPERIMENT 1a: SPEECH RATE

Figure B2 illustrates the mean percentage of times listeners heard the manipulated versions of the voices as sounding the same as the original (i.e., unmanipulated) versions before the data was collapsed across plus and minus manipulations, for speech rate.

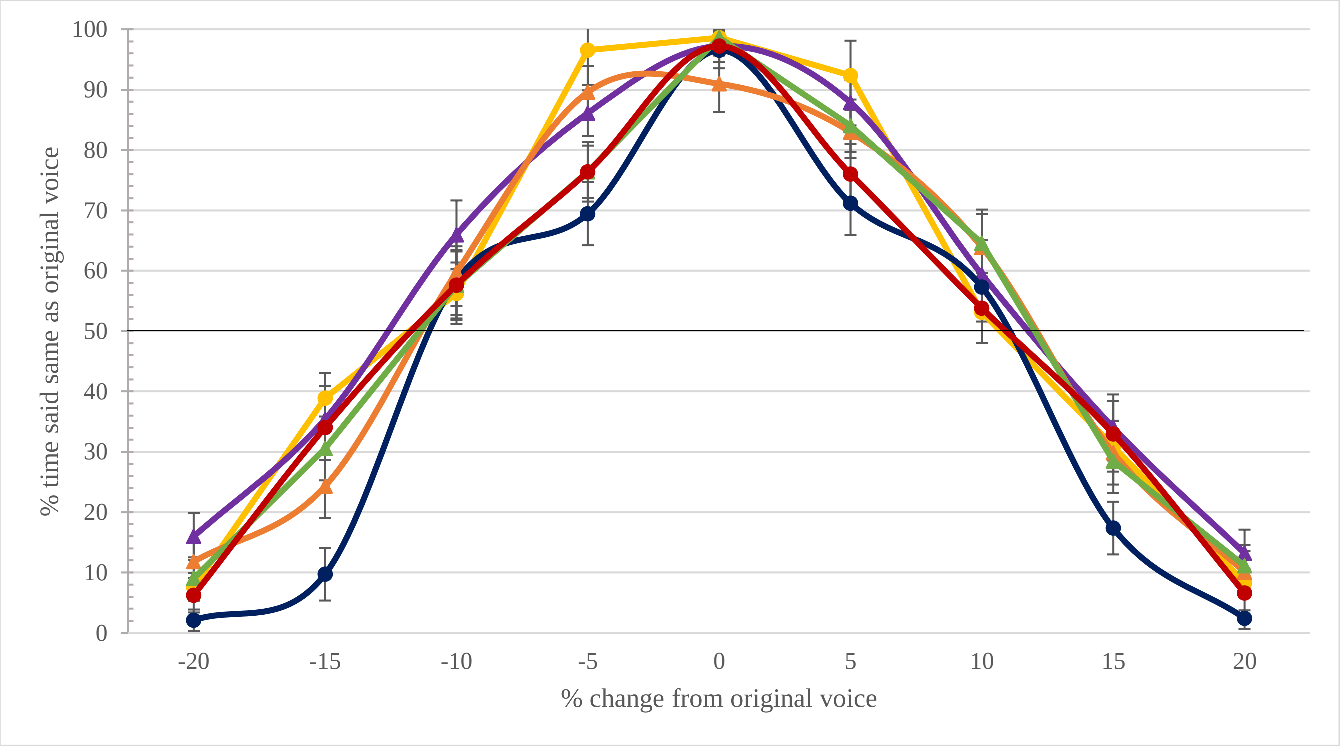


Figure B2: Line graph depicting percentage of times listeners heard voice as sounding the 'same' as the original voice, for speech rate. Each of the six voices are depicted by a different line colour. Male voices are depicted using the triangle symbol and female voices are depicted using the circle symbol. 95% confidence intervals are also shown.

APPENDIX C1. ADDITIONAL FIGURES FOR EXPERIMENT 2: FUNDAMENTAL FREQUENCY (F0)

Figure C1 illustrates the mean percentage of times listeners heard the manipulated versions of the voices as sounding like the same speaker (i.e., same identity) as the original (i.e., unmanipulated) versions before the data was collapsed across plus and minus manipulations, for F0.

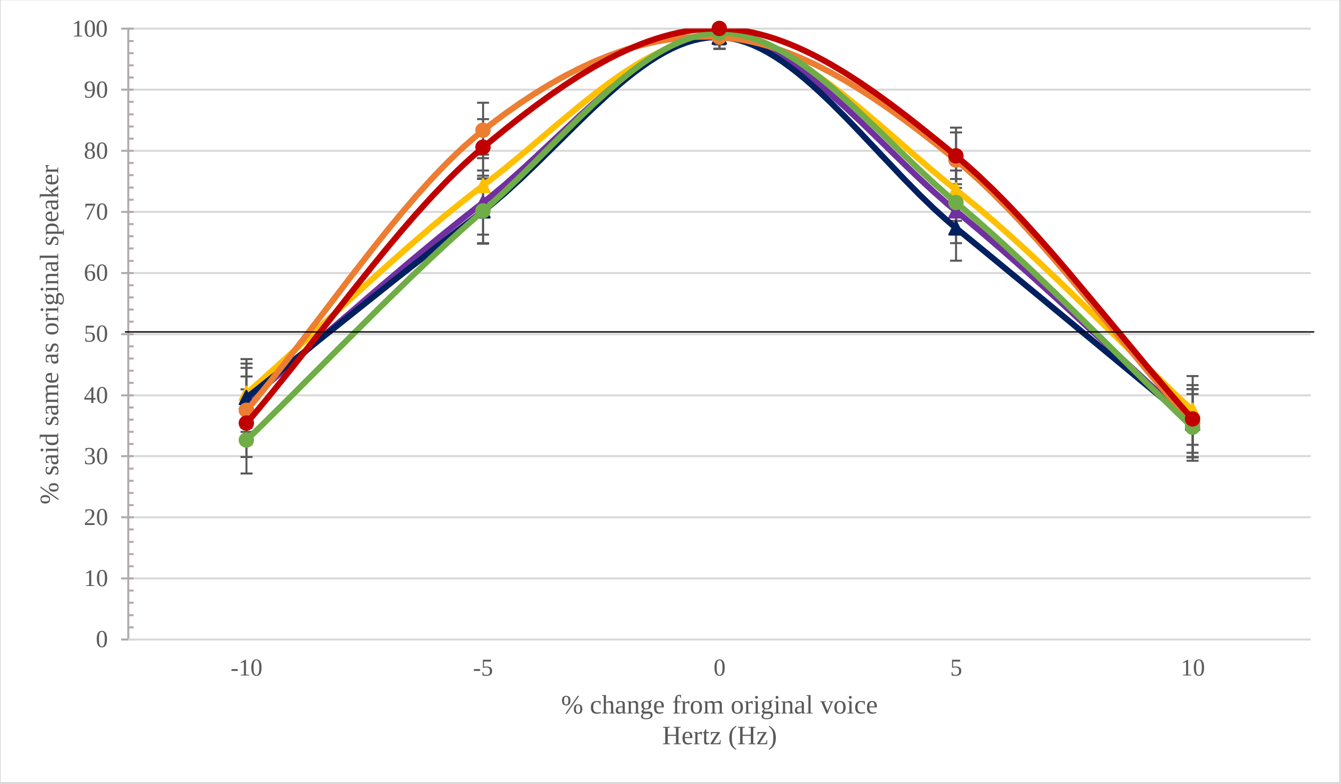


Figure C1: Line graph depicting percentage of times listeners heard voice as sounding the same speaker (i.e., the same identity) as the original voice, for F0. Each of the six voices are depicted by a different line colour. Male voices are depicted using the triangle symbol and female voices are depicted using the circle symbol. 95% confidence intervals are also shown.

APPENDIX C2. ADDITIONAL FIGURES FOR EXPERIMENT 2: SPEECH RATE

Figure C2 illustrates the mean percentage of times listeners heard the manipulated versions of the voices as sounding like the same speaker (i.e., same identity) as the original (i.e., unmanipulated) versions before the data was collapsed across plus and minus manipulations, for speech rate.

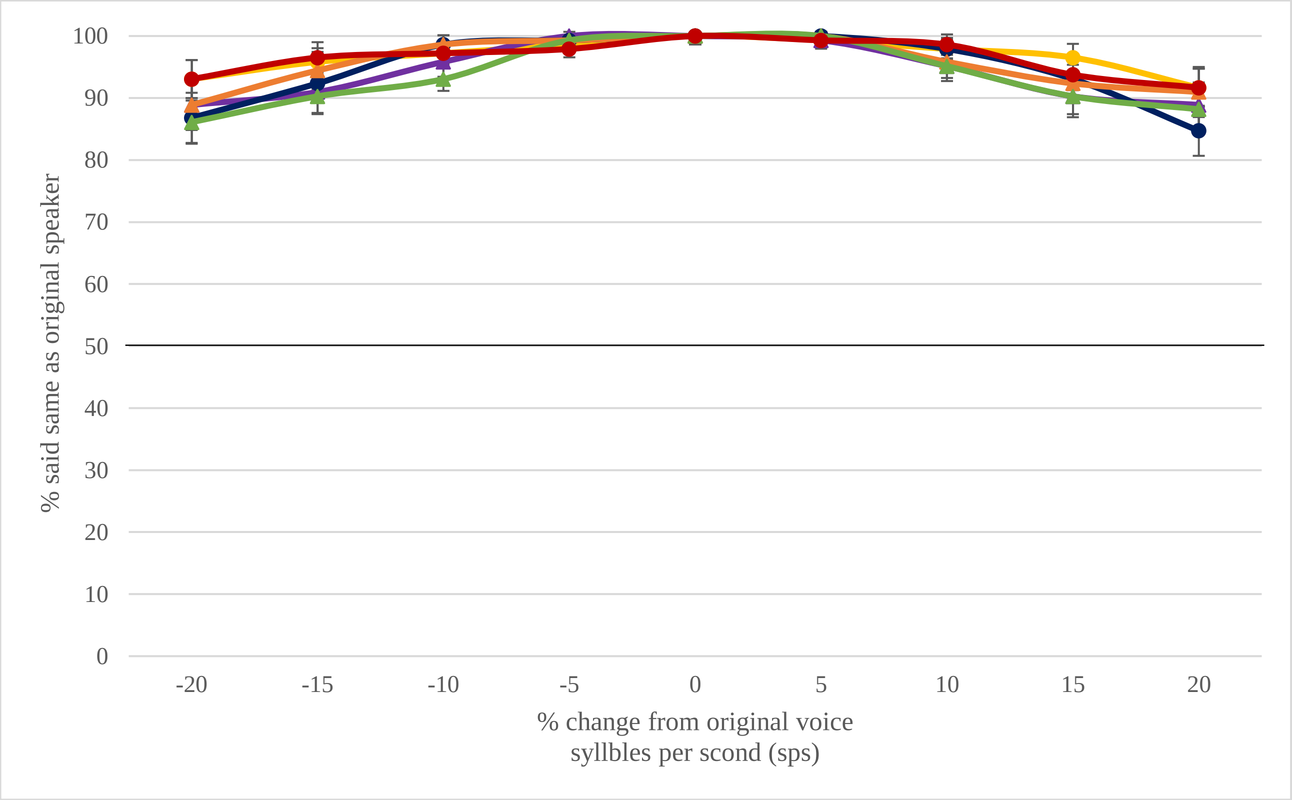


Figure C2: Line graph depicting percentage of times listeners heard voice as sounding the same speaker (i.e., the same identity) as the original voice, for speech rate. Each of the six voices are depicted by a different line colour. Male voices are depicted using the triangle symbol and female voices are depicted using the circle symbol. 95% confidence intervals are also shown.

APPENDIX D1. OUTCOME OF THE 5-WAY ANOVA FOR FUNDAMENTAL FREQUENCY (F0): EXPERIMENT 5a

Table D1 displays the degrees of freedom (*df*), F ratios (*F*), effect sizes (generalised eta squared; η_g^2), and adjusted *p* values using the Hochberg correction (*p*) for all the main study variables and associated interactions in Experiment 5a, for F0.

Table D1: *Outcome of the 5-way ANOVA for Experiment 5a with degrees of freedom (df), F ratios (F), effect sizes (η_g^2), and adjusted p values (p) using the Hochberg correction.*

	df	*F*	η_g^2	*p*
Listener Sex	1 (28)	2.69	.009	.93
Sex of Voice	1 (28)	8.16	.005	.22
Target F0	2 (56)	3.08	.003	.93
Direction of Manipulation	1 (28)	94.56	.07	.03*
Magnitude of Distractor Change	2 (56)	50.75	.13	.03*
Sex of Voice x Listener Sex	1 (28)	.01	.000004	.93
Target F0 x Listener Sex	2 (56)	.69	.003	.93
Direction of Manipulation x Listener Sex	1 (28)	3.28	.007	.58
Magnitude of Distractor Change x Listener Sex	2 (56)	.11	.0003	.93
Sex of Voice x Target F0	2 (56)	.66	.0005	.93
Sex of Voice x Direction of Manipulation	1 (28)	5.58	.002	.58
Target F0 x Direction of Manipulation	2 (56)	9.27	.03	.03*

Sex of Voice x Magnitude of Distractor Change	2 (56)	1.35	.000004	.93
Target F0 x Magnitude of Distractor Change	4 (112)	3.36	.002	.31
Direction of Manipulation x Magnitude of Distractor Change	2 (56)	18.01	.04	.03*
Sex of Voice x Target F0 x Listener Sex	2 (56)	4.14	.007	.53
Sex of Voice x Direction of Manipulation x Listener Sex	1 (28)	.39	.001	.93
Target F0 x Direction of Manipulation x Listener Sex	2 (56)	.74	.008	.93
Sex of Voice x Target F0 x Direction of Manipulation	2 (56)	2.92	.001	.93
Sex of Voice x Magnitude of Distractor Change x Listener Sex	2 (56)	1.30	.002	.93
Target F0 x Magnitude of Distractor Change x Listener Sex	4 (112)	1.52	.002	.93
Sex of Voice x Target F0 x Magnitude of Distractor Change	4 (112)	1.69	.005	.93
Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	2 (56)	1.15	.00009	.93
Sex of Voice x Direction of Manipulation x Magnitude of Distractor Change	2 (56)	.42	.00009	.93
Target F0 x Direction of Manipulation x Magnitude of Distractor Change	4 (112)	1.73	.009	.93

Sex of Voice x Target F0 x Direction of Manipulation x Listener Sex	2 (56)	.07	.0008	.93
Sex of Voice x Target F0 x Magnitude of Distractor Change x Listener Sex	2 (56)	1.45	.005	.93
Sex of Voice x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	2 (56)	.48	.00009	.93
Target F0 x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	4 (112)	3.26	.01	.58
Sex of Voice x Target F0 x Direction of Manipulation x Magnitude of Distractor Change	4 (112)	3.42	.01	.93
Sex of Voice x Target F0 x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	4 (112)	.68	.0009	.93

Note. Error degrees of freedom (df) are shown in parentheses. * denote significance at the p<0.05 level.

APPENDIX D2. OUTCOME OF THE 5-WAY ANOVA FOR SPEECH RATE: EXPERIMENT 5b

Table D2 displays the degrees of freedom (*df*), F ratios (*F*), effect sizes (generalised eta squared; η_g^2), and adjusted *p* values using the Hochberg correction (*p*) for all the main study variables and associated interactions in Experiment 5b, for speech rate.

Table D2: *Outcome of the 5-way ANOVA for Experiment 5b, with degrees of freedom (df), F ratios (F), effect sizes (η_g^2), and adjusted p values (p) using the Hochberg correction.*

	df	*F*	η_g^2	*p*
Listener Sex	1 (28)	.003	.00005	.996
Sex of Voice	1 (28)	.11	.0006	.996
Target Speech Rate	2 (56)	2.82	.006	.996
Direction of Manipulation	1 (28)	12.55	.02	.03*
Magnitude of Distractor Change	2 (56)	50.27	.10	.03*
Sex of Voice x Listener Sex	1 (28)	.002	.00005	.996
Target Speech Rate x Listener Sex	2 (56)	.34	.001	.996
Direction of Manipulation x Listener Sex	1 (28)	.021	.00002	.996
Magnitude of Distractor Change x Listener Sex	2 (56)	.004	.000001	.996
Sex of Voice x Target Speech Rate	2 (56)	2.39	.006	.996
Sex of Voice x Direction of Manipulation	1 (28)	.027	.00005	.996
Target Speech Rate x Direction of Manipulation	2 (56)	15.13	.06	.03*
Sex of Voice x Magnitude of Distractor Change	2 (56)	3.27	.001	.996

Target Speech Rate x Magnitude of Distractor Change	4 (112)	1.44	.004	.996
Direction of Manipulation x Magnitude of Distractor Change	2 (56)	.65	.00004	.996
Sex of Voice x Target Speech Rate x Listener Sex	2 (56)	.58	.0002	.996
Sex of Voice x Direction of Manipulation x Listener Sex	1 (28)	.011	.0005	.996
Target Speech Rate x Direction of Manipulation x Listener Sex	2 (56)	1.05	.004	.996
Sex of Voice x Target Speech Rate x Direction of Manipulation	2 (56)	.24	.00003	.996
Sex of Voice x Magnitude of Distractor Change x Listener Sex	2 (56)	.37	.0004	.996
Target Speech Rate x Magnitude of Distractor Change x Listener Sex	4 (112)	2.01	.01	.996
Sex of Voice x Target Speech Rate x Magnitude of Distractor Change	4 (112)	3.17	.005	.43
Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	2 (56)	.60	.0008	.996
Sex of Voice x Direction of Manipulation x Magnitude of Distractor Change	2 (56)	.04	.00005	.996
Target Speech Rate x Direction of Manipulation x Magnitude of Distractor Change	4 (112)	1.89	.0001	.996
Sex of Voice x Target Speech Rate x Direction of Manipulation x Listener Sex	2 (56)	.14	.00008	.996
Sex of Voice x Target Speech Rate x Magnitude of Distractor Change x Listener Sex	2 (56)	.26	.0001	.996
Sex of Voice x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	2 (56)	1.53	.002	.996

Target Speech Rate x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	4 (112)	.102	.0002	.996
Sex of Voice x Target Speech Rate x Direction of Manipulation x Magnitude of Distractor Change	4 (112)	.45	.0009	.996
Sex of Voice x Target Speech Rate x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	4 (112)	.22	.0003	.996

Note. Error degrees of freedom (df) are shown in parentheses. * denote significance at the p<0.05 level.

APPENDIX E. OUTCOME OF THE 5-WAY ANOVA FOR FUNDAMENTAL FREQUENCY (F0): EXPERIMENT 6

Table E displays the degrees of freedom (*df*), F ratios (*F*), effect sizes (generalised eta squared; η_g^2), and adjusted *p* values using the Hochberg correction (*p*) for all the main study variables and associated interactions in Experiment 6, for F0.

Table E: *Outcome of the 5-way ANOVA for Experiment 6, with degrees of freedom (df), F ratios (F), effect sizes (η_g^2), and adjusted p values (p) using the Hochberg correction.*

	df	*F*	η_g^2	*p*
Listener Sex	1 (28)	1.05	.0008	.96
Sex of Voice	1 (28)	11.58	.006	.052
Target F0	2 (56)	2.72	.003	.96
Direction of Manipulation	1 (28)	27.90	.06	.03*
Magnitude of Distractor Change	1 (28)	12.20	.009	.05
Sex of Voice x Listener Sex	1 (28)	1.37	.0007	.96
Target F0 x Listener Sex	2 (56)	.26	.0003	.96
Direction of Manipulation x Listener Sex	1 (28)	1.76	.00007	.96
Magnitude of Distractor Change x Listener Sex	1 (28)	4.08	.003	.96
Sex of Voice x Target F0	2 (56)	1.40	.005	.96
Sex of Voice x Direction of Manipulation	1 (28)	.12	.00006	.96
Target F0 x Direction of Manipulation	2 (56)	11.79	.02	.03*
Sex of Voice x Magnitude of Distractor Change	1 (28)	.003	.002	.96
Target F0 x Magnitude of Distractor Change	2 (56)	2.31	.007	.96
Direction of Manipulation x Magnitude of Distractor Change	1 (28)	18.10	.02	.03*
Sex of Voice x Target F0 x Listener Sex	2 (56)	3.23	.003	.96
Sex of Voice x Direction of Manipulation x Listener Sex	1 (28)	.37	.001	.96

Target F0 x Direction of Manipulation x Listener Sex	2 (56)	.20	.0008	.96
Sex of Voice x Target F0 x Direction of Manipulation	2 (56)	.54	.0008	.96
Sex of Voice x Magnitude of Distractor Change x Listener Sex	1 (28)	2.72	.009	.96
Target F0 x Magnitude of Distractor Change x Listener Sex	2 (56)	4.53	.005	.38
Sex of Voice x Target F0 x Magnitude of Distractor Change	2 (56)	.06	.0008	.96
Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	1 (28)	.003	.00009	.96
Sex of Voice x Direction of Manipulation x Magnitude of Distractor Change	1 (28)	4.19	.007	.96
Target F0 x Direction of Manipulation x Magnitude of Distractor Change	1.52 (42.58)	.45	.0008	.96
Sex of Voice x Target F0 x Direction of Manipulation x Listener Sex	2 (56)	.24	.008	.96
Sex of Voice x Target F0 x Magnitude of Distractor Change x Listener Sex	2 (56)	2.50	.009	.96
Sex of Voice x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	1 (28)	.77	.0007	.96
Target F0 x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	2 (56)	3.30	.01	.96
Sex of Voice x Target F0 x Direction of Manipulation x Magnitude of Distractor Change	2 (56)	.43	.0001	.96
Sex of Voice x Target F0 x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	2 (56)	.38	.0009	.96

Note. Error degrees of freedom (df) are shown in parentheses. * denote significance at the p<0.05 level.

APPENDIX F. OUTCOME OF THE 5-WAY ANOVA FOR FUNDAMENTAL FREQUENCY (F0): EXPERIMENT 7

Table F displays the degrees of freedom (*df*), F ratios (*F*), effect sizes (generalised eta squared; η_g^2), and adjusted *p* values using the Hochberg correction (*p*) for all the main study variables and associated interactions in Experiment 7, for F0.

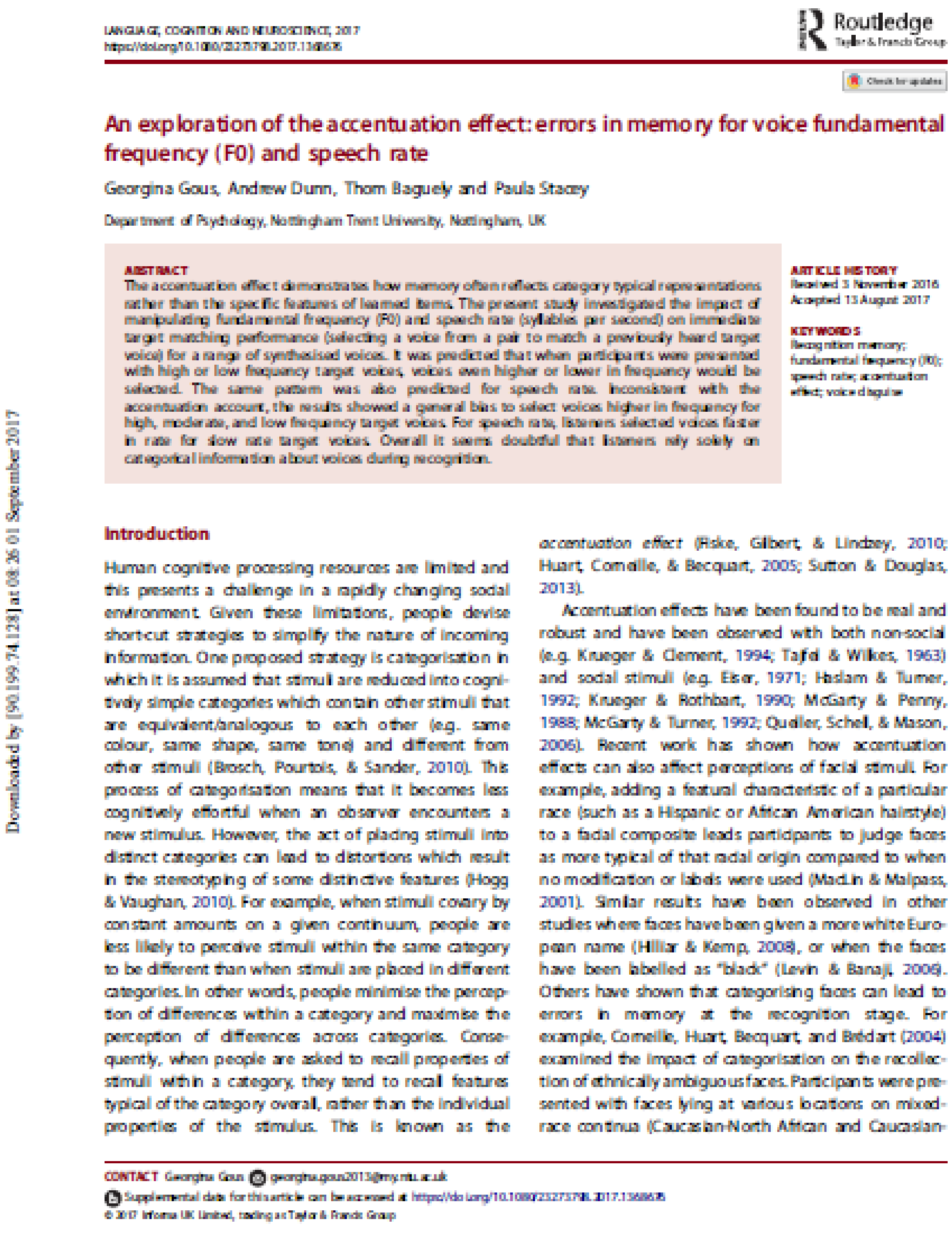
Table F: *Outcome of the 5-way ANOVA for Experiment 7, with degrees of freedom (df), F ratios (F), effect sizes (η_g^2), and adjusted p values (p) using the Hochberg correction.*

	df	*F*	η_g^2	*p*
Listener Sex	1 (28)	2.51	.009	.93
Sex of Voice	1 (28)	4.65	.0005	.93
Target F0	2 (56)	2.92	.001	.93
Direction of Manipulation	1 (28)	40.84	.08	.027*
Magnitude of Distractor Change	1 (28)	49.60	.04	.027*
Sex of Voice x Listener Sex	1 (28)	1.16	.0008	.93
Target F0 x Listener Sex	2 (56)	4.45	.01	.40
Direction of Manipulation x Listener Sex	1 (28)	.91	.0007	.93
Magnitude of Distractor Change x Listener Sex	1 (28)	2.53	.003	.93
Sex of Voice x Target F0	2 (56)	3.35	.009	.93
Sex of Voice x Direction of Manipulation	1 (28)	21.34	.03	.027*
Target F0 x Direction of Manipulation	2 (56)	26.54	.09	.027*
Sex of Voice x Magnitude of Distractor Change	1 (28)	1.21	.00005	.93
Target F0 x Magnitude of Distractor Change	2 (56)	.59	.0002	.93
Direction of Manipulation x Magnitude of Distractor Change	1 (28)	9.92	.007	.10
Sex of Voice x Target F0 x Listener Sex	2 (56)	.49	.0007	.93
Sex of Voice x Direction of Manipulation x Listener Sex	1 (28)	.68	.001	.93

Target F0 x Direction of Manipulation x Listener Sex	2 (56)	.32	.009	.93
Sex of Voice x Target F0 x Direction of Manipulation	2 (56)	2.29	.001	.93
Sex of Voice x Magnitude of Distractor Change x Listener Sex	1 (28)	.30	.00003	.93
Target F0 x Magnitude of Distractor Change x Listener Sex	2 (56)	.79	.007	.93
Sex of Voice x Target F0 x Magnitude of Distractor Change	2 (56)	.77	.008	.93
Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	1 (28)	1.10	.00009	.93
Sex of Voice x Direction of Manipulation x Magnitude of Distractor Change	1 (28)	.17	.0008	.93
Target F0 x Direction of Manipulation x Magnitude of Distractor Change	1.52 (42.58)	.81	.0007	.93
Sex of Voice x Target F0 x Direction of Manipulation x Listener Sex	2 (56)	1.21	.009	.93
Sex of Voice x Target F0 x Magnitude of Distractor Change x Listener Sex	2 (56)	.30	.0008	.93
Sex of Voice x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	1 (28)	.37	.00009	.93
Target F0 x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	2 (56)	.73	.00008	.93
Sex of Voice x Target F0 x Direction of Manipulation x Magnitude of Distractor Change	2 (56)	.36	.00008	.93
Sex of Voice x Target F0 x Direction of Manipulation x Magnitude of Distractor Change x Listener Sex	2 (56)	.07	.00009	.93

Note. Error degrees of freedom (df) are shown in parentheses. * denote significance at the p<0.05 level.

APPENDIX G. PUBLISHED ARTICLE (Gous, Dunn, Baguley, & Stacey, 2017)



LANGUAGE, COGNITION AND NEUROSCIENCE, 2017
https://doi.org/10.1080/23273798.2017.1368678

Routledge
Taylor & Francis Group

Check for updates

An exploration of the accentuation effect: errors in memory for voice fundamental frequency (F0) and speech rate

Georgina Gous, Andrew Dunn, Thom Baguely and Paula Stacey

Department of Psychology, Nottingham Trent University, Nottingham, UK

ABSTRACT
The accentuation effect demonstrates how memory often reflects category typical representations rather than the specific features of learned items. The present study investigated the impact of manipulating fundamental frequency (F0) and speech rate (syllables per second) on immediate target matching performance (selecting a voice from a pair to match a previously heard target voice) for a range of synthesised voices. It was predicted that when participants were presented with high or low frequency target voices, voices even higher or lower in frequency would be selected. The same pattern was also predicted for speech rate. Inconsistent with the accentuation account, the results showed a general bias to select voices higher in frequency for high, moderate, and low frequency target voices. For speech rate, listeners selected voices faster in rate for slow rate target voices. Overall it seems doubtful that listeners rely solely on categorical information about voices during recognition.

ARTICLE HISTORY
Received 3 November 2016
Accepted 13 August 2017

KEYWORDS
Recognition memory; fundamental frequency (F0); speech rate; accentuation effect; voice disguise

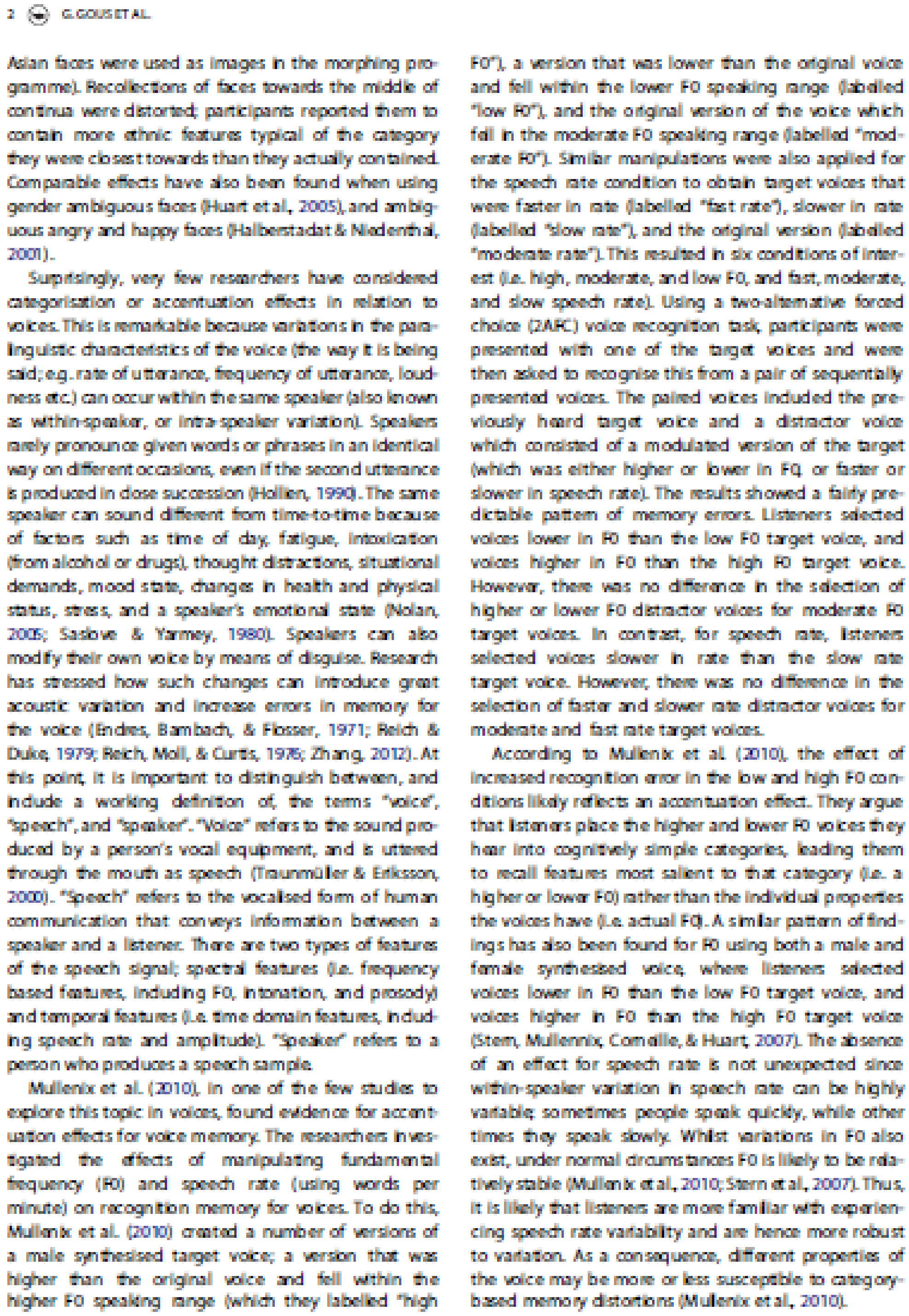
Introduction

Human cognitive processing resources are limited and this presents a challenge in a rapidly changing social environment. Given these limitations, people devise short-cut strategies to simplify the nature of incoming information. One proposed strategy is categorisation in which it is assumed that stimuli are reduced into cognitively simple categories which contain other stimuli that are equivalent/analogous to each other (e.g. same colour, same shape, same tone) and different from other stimuli (Brosch, Pourtois, & Sander, 2010). This process of categorisation means that it becomes less cognitively effortful when an observer encounters a new stimulus. However, the act of placing stimuli into distinct categories can lead to distortions which result in the stereotyping of some distinctive features (Hogg & Vaughan, 2010). For example, when stimuli covary by constant amounts on a given continuum, people are less likely to perceive stimuli within the same category to be different than when stimuli are placed in different categories. In other words, people minimise the perception of differences within a category and maximise the perception of differences across categories. Consequently, when people are asked to recall properties of stimuli within a category, they tend to recall features typical of the category overall, rather than the individual properties of the stimulus. This is known as the *accentuation effect* (Fiske, Gilbert, & Lindzey, 2010; Huart, Corneille, & Becquart, 2005; Sutton & Douglas, 2013).

Accentuation effects have been found to be real and robust and have been observed with both non-social (e.g. Krueger & Clement, 1994; Tajfel & Wilkes, 1963) and social stimuli (e.g. Eiser, 1971; Haslam & Turner, 1992; Krueger & Rothbart, 1990; McGarty & Penny, 1988; McGarty & Turner, 1992; Queller, Schell, & Mason, 2006). Recent work has shown how accentuation effects can also affect perceptions of facial stimuli. For example, adding a featural characteristic of a particular race (such as a Hispanic or African American hairstyle) to a facial composite leads participants to judge faces as more typical of that racial origin compared to when no modification or labels were used (MacLin & Malpass, 2001). Similar results have been observed in other studies where faces have been given a more white European name (Hilliar & Kemp, 2008), or when the faces have been labelled as "black" (Levin & Banaji, 2006). Others have shown that categorising faces can lead to errors in memory at the recognition stage. For example, Corneille, Huart, Becquart, and Brédart (2004) examined the impact of categorisation on the recollection of ethnically ambiguous faces. Participants were presented with faces lying at various locations on mixed-race continua (Caucasian-North African and Caucasian-

CONTACT Georgina Gous georgina.gous2013@my.ntu.ac.uk

Supplemental data for this article can be accessed at https://doi.org/10.1080/23273798.2017.1368678



Asian faces were used as images in the morphing programme). Recollections of faces towards the middle of continua were distorted; participants reported them to contain more ethnic features typical of the category they were closest towards than they actually contained. Comparable effects have also been found when using gender ambiguous faces (Huart et al., 2005), and ambiguous angry and happy faces (Halberstadat & Niedenthal, 2001).

Surprisingly, very few researchers have considered categorisation or accentuation effects in relation to voices. This is remarkable because variations in the paralinguistic characteristics of the voice (the way it is being said; e.g. rate of utterance, frequency of utterance, loudness etc.) can occur within the same speaker (also known as within-speaker, or intra-speaker variation). Speakers rarely pronounce given words or phrases in an identical way on different occasions, even if the second utterance is produced in close succession (Hollien, 1990). The same speaker can sound different from time-to-time because of factors such as time of day, fatigue, intoxication (from alcohol or drugs), thought distractions, situational demands, mood state, changes in health and physical status, stress, and a speaker's emotional state (Nolan, 2005; Saslove & Yarmey, 1980). Speakers can also modify their own voice by means of disguise. Research has stressed how such changes can introduce great acoustic variation and increase errors in memory for the voice (Endres, Bambach, & Flosser, 1971; Reich & Duke, 1979; Reich, Moll, & Curtis, 1976; Zhang, 2012). At this point, it is important to distinguish between, and include a working definition of, the terms "voice", "speech", and "speaker". "Voice" refers to the sound produced by a person's vocal equipment, and is uttered through the mouth as speech (Traunmüller & Eriksson, 2000). "Speech" refers to the vocalised form of human communication that conveys information between a speaker and a listener. There are two types of features of the speech signal; spectral features (i.e. frequency based features, including F0, intonation, and prosody) and temporal features (i.e. time domain features, including speech rate and amplitude). "Speaker" refers to a person who produces a speech sample.

Mullenix et al. (2010), in one of the few studies to explore this topic in voices, found evidence for accentuation effects for voice memory. The researchers investigated the effects of manipulating fundamental frequency (F0) and speech rate (using words per minute) on recognition memory for voices. To do this, Mullenix et al. (2010) created a number of versions of a male synthesised target voice; a version that was higher than the original voice and fell within the higher F0 speaking range (which they labelled "high F0"), a version that was lower than the original voice and fell within the lower F0 speaking range (labelled "low F0"), and the original version of the voice which fell in the moderate F0 speaking range (labelled "moderate F0"). Similar manipulations were also applied for the speech rate condition to obtain target voices that were faster in rate (labelled "fast rate"), slower in rate (labelled "slow rate"), and the original version (labelled "moderate rate"). This resulted in six conditions of interest (i.e. high, moderate, and low F0, and fast, moderate, and slow speech rate). Using a two-alternative forced choice (2AFC) voice recognition task, participants were presented with one of the target voices and were then asked to recognise this from a pair of sequentially presented voices. The paired voices included the previously heard target voice and a distractor voice which consisted of a modulated version of the target (which was either higher or lower in F0, or faster or slower in speech rate). The results showed a fairly predictable pattern of memory errors. Listeners selected voices lower in F0 than the low F0 target voice, and voices higher in F0 than the high F0 target voice. However, there was no difference in the selection of higher or lower F0 distractor voices for moderate F0 target voices. In contrast, for speech rate, listeners selected voices slower in rate than the slow rate target voice. However, there was no difference in the selection of faster and slower rate distractor voices for moderate and fast rate target voices.

According to Mullenix et al. (2010), the effect of increased recognition error in the low and high F0 conditions likely reflects an accentuation effect. They argue that listeners place the higher and lower F0 voices they hear into cognitively simple categories, leading them to recall features most salient to that category (i.e. a higher or lower F0) rather than the individual properties the voices have (i.e. actual F0). A similar pattern of findings has also been found for F0 using both a male and female synthesised voice, where listeners selected voices lower in F0 than the low F0 target voice, and voices higher in F0 than the high F0 target voice (Stern, Mullennix, Corneille, & Huart, 2007). The absence of an effect for speech rate is not unexpected since within-speaker variation in speech rate can be highly variable; sometimes people speak quickly, while other times they speak slowly. Whilst variations in F0 also exist, under normal circumstances F0 is likely to be relatively stable (Mullenix et al., 2010; Stern et al., 2007). Thus, it is likely that listeners are more familiar with experiencing speech rate variability and are hence more robust to variation. As a consequence, different properties of the voice may be more or less susceptible to category-based memory distortions (Mullenix et al., 2010).

The present study examined the impact of variations in F0 and speech rate for a set of unfamiliar synthesised voices in a similar manner to Mullenix et al. (2010), but with a number of important extensions and modifications to the procedure. First, we used a slightly larger set of synthesised voices (two male, two female), which increases the generalisability of the findings. Second, we kept the target and distractor voices within a F0 and speech rate range that is typical in the population for English speakers. This is important given that it is highly unusual to hear voices outside of the typical male and female range in everyday situations. Third, we also included sex of voice and listener sex as independent variables in our design. This is important given that research has emphasised sex differences in verbal episodic memory tasks, with women often performing at a higher level than men (Herlitz, Nilsson, & Backman, 1997; Lewin, Wolgers, & Herlitz, 2001; McGivern et al., 1997). Others have also reported an own-gender bias (i.e. better recognition performance for voices of an observers own sex) for unfamiliar voices (Roebuck & Wilding, 1993).

Following Mullenix et al. (2010), we investigated the impact of manipulating overall mean F0 (in Hz) and speech rate (in syllables per second[1]) on immediate target matching performance. We used a 2AFC procedure in which listeners were asked to recognise a target voice from a voice pair that contained the previously heard target voice and a modulated version of the voice. There were six conditions of interest (i.e. high, moderate, and low F0, and fast, moderate, and slow speech rate). In keeping with the terminology used by Mullenix et al. (2010), three versions for each target voice were created for both the F0 and speech rate conditions. For the F0 condition, we created a version that was higher than the original voice and fell within the higher F0 speaking range (labelled "high F0"), a version that was lower than the original voice and fell within the lower F0 speaking range (labelled "low F0"), and the original version of the voice which fell in the moderate F0 speaking range (labelled "moderate F0"). Similarly, for the speech rate condition, we created a version that was faster than the original voice (labelled "fast rate"), a version that was slower than the original voice (labelled "slow rate"), and the original version of the voice (labelled "moderate rate"). To obtain our distractor voices, we further increased and decreased each target voice in F0 and speech rate.

It was expected that the results would parallel those of Mullenix et al. (2010). For F0, we predicted that there would be a memory bias for high and low F0 target voices but not for moderate F0 target voices. Specifically, we expected to see an increase in the selection of voices higher in F0 when high F0 target voices were presented, and an increase in the selection of voices lower in F0 when low F0 target voices were presented. We were more tentative with our predictions for speech rate since Mullenix et al. (2010) found no memory biases for their speech rate manipulations, but in line with the accentuation effect, we hypothesised that people would be more likely to select distractors that were faster in rate for voices that had a fast speech rate, and to select distractors slower in rate for voices that had a slow speech rate.

Method

Design

The participants were arbitarily allocated to either the F0 condition or the speech rate condition. For each condition, the experiment employed a $2 \times 2 \times 3 \times 3 \times 2$ mixed factorial design. The between-subjects factor was listener sex (male or female). The within-subjects factors were sex of voice (male or female), target type (*for F0:* high, moderate or low, *for speech rate:* fast, moderate or slow), magnitude of distractor change (*for F0:* 5%, 7%, or 10%, *for speech rate:* 10%, 12%, or 20%) and direction of manipulation (*for F0:* increase or decrease in F0, *for speech rate:* increase or decrease in rate). The dependent variable measured was mean percentage of errors made (i.e. percentage of time listeners choose the distractor voice instead of the target voice).

Participants

A total of 60 undergraduate students (30 males; 30 females) were recruited from Nottingham Trent University and they received course credit for their participation. The inclusion criteria for the study required individuals to be between 18 and 30 years of age, have no known hearing deficits, have English as their first language, and not undergone any musical training.

A total of 30 individuals contributed to the F0 condition (15 males; 15 females). The ages of the participants ranged from 18 to 27 years old (M = 21.03 years, SD = 2.09 years). A further 30 individuals contributed to the speech rate condition (15 males; 15 females). The ages of the participants ranged from 18 to 30 years old (M = 21.72 years, SD = 2.62 years).

Stimuli and materials

Natural Reader 12.0 (http://www.naturalreaders.com/index.html) was used to create the four voice samples (four different identities, two male and two female).

Natural Reader is a text-to-speech software with realistic and natural sounding synthesised voices, generating speech samples from concatenated pieces of real human speech. Synthetic speech was used because of the need for precisely controlled stimuli that varied in F0 and speech rate. Concatenated speech also gives the advantage of sounding more natural than fully synthesised speech. The target speech samples were created by typing the following phrase "*Spring is the season where flowers appear, summer is the warmest season of the year.*" in Natural Reader. The four original voice samples were then manipulated in F0 and speech rate using Audacity® software (http://www.audacityteam.org/).[2] Audacity® is a free audio software that can be used to edit sounds and was chosen to manipulate the voices because it allowed us to alter one characteristic (F0 and speech rate) whilst holding the other constant.

In order to select stimuli for the main experiment, we pilot tested a number of speech samples for both the F0 and speech rate conditions. This enabled us to create additional voice samples that were both higher and lower in F0, and faster and slower in speech rate. Using a perceptual discrimination paradigm, 72 participants (36 males and 36 females) were given a 2AFC (same/different key press) voice matching task. Participants responded by indicating whether the two stimuli on each trial were the "same" or "different". The stimuli were presented as within voice pairs with a 1 s inter-stimulus separating them. The original target voice was used as the "standard" stimulus and presented on all trials. The standard stimulus was paired with either itself or a modulated version (increased/decreased in F0, or increased/decreased in speech rate) and presented in a random order. For F0 the modulated versions were increased and decreased by 5% and 10%, and for speech rate they were increased and decreased by 5%, 10%, 15%, and 20%. This was considered appropriate given that a modification in F0 elicited a greater audible change than it did for speech rate. This resulted in a total of 104 trials (13 trials for each voice, with each trial being counterbalanced and presented twice). For F0, a setting of plus and minus 6.63% was judged as 50% discriminable, and for speech rate this was 11.52%. Smaller manipulations in F0 and speech rate were judged as sounding more similar to the target voice, whereas larger manipulations were judged as sounding less similar to the target voice. The pilot testing also allowed us to determine whether the distractor voices chosen for the main experiment were discriminable from the target voice.

For each of the original synthesised voices, we used the 10% modulated versions to obtain target voices in the higher and lower F0 range, and the 20% modulated versions to obtain target voices in the faster and slower speech rate range. For F0, the typical adult male will have an F0 between 80 and 180 Hz, and for an adult female this will be between 165 and 255 Hz (Titze, 1994). For speech rate, the typical range for male and female speech is 3.3 to 5.9 syllables/s (Arnfield, Roach, Setter, Greasley, & Horton, 1995). It is important to emphasise however that different speaking styles typically entail different speaking rates and therefore absolute values can differ (Brown, 2014). All four original (unedited) voice samples fell within the moderate speaking range for both F0 and speech rate, and thus acted as moderate target voices. This resulted in six experimental conditions of interest; low F0, moderate F0 and high F0, and slow speech rate, moderate speech rate, and fast speech rate.

Based on the findings from the pilot study, we increased and decreased each target voice by a further 5%, 7%, and 10% for F0, and by a further 10%, 12%, and 20% for speech rate, to obtain our distractor speech samples. This resulted in a total of six modulated versions (i.e. distractor voices) for each target voice sample; three increased in F0 or speech rate, and three decreased in F0 or speech rate (refer to Appendix B; Table B1 for F0 values, and Table B2 for speech rate values). All of the voices samples, whether targets or distractors, fell within the typical F0 and speech rate range for normally voiced speech for English speakers. The distractor voices spoke the same phrase as the target voices.

The voice samples were tested to determine how naturalistic (i.e. how realistic, or lifelike) they sounded (refer to Appendix C for further details). This is important because the authors wanted to ensure that the voices used were generalisable to those voices that are heard in a real-world environment. Mean naturalness ratings across all of the voices were 73.46% for F0 manipulations and 72.6% for speech rate manipulations. Whilst we recognise that these are not perfect, these values are nevertheless slightly higher than those identified elsewhere (e.g. 70%) (see Jreige, Patel, & Bunnell, 2009), and are a reasonable indication that the synthesised voices used for experimentation are representative of real voices. It should also be noted that the voice samples contained smooth formant transitions and there were no intonational irregularities or prosodic mismatches across words.

The voice samples were also tested to determine whether the four voices used for experimentation were perceived as being different speakers (i.e. different identities). This was important because the authors wanted to ensure that all of the voices used for experimentation were distinct from each and that they would not be confused with another voice that they had previously heard (refer to Appendix D for further details). The results showed that listeners could correctly determine that the voices were different speakers with almost 100% accuracy. Thus, it can be assumed that the voices used for experimentation were distinct from each other and perceived as being different speakers.

All of the speech samples were saved as separate.wav files and presented binaurally using Sony dynamic stereo headphones (Model No. MDR-V150). The experiment was run on a Sony Vaio laptop computer (Model No. SVF153B1YM) using PsychoPy version 1.7701 (Peirce, 2007) to control the presentation and collect participant responses.

Procedure

The participants were arbitrarily allocated to either the F0 or speech rate condition. For each experimental condition, there were 144 trials in total. Specifically, there were four different voices (two male and two female), each with three target voices (high, moderate, and low F0, or fast, moderate, and slow speech rate). For each target voice, there were 12 trials in total (each of the three targetvoices were paired with one of the six distractor voices, with each trial being presented twice). In each trial, participants were first presented with one of the target speech samples. After a one second gap, the participants were presented with sequentially paired voices that included the target voice (present in all trials) and one of the six distractor voices (that was either increased or decreased in F0 or speech rate).There was a one second inter-stimulus interval between presentation of each voice. The trials were counterbalanced so that half the time the target voice was presented first, and half the time the target voice was presented second. The order of the trials were randomised across participants using PsychoPy. Following presentation of each trial the participants were asked "which voice matched the voice you previously heard, voice one or voice two?". The participants had to indicate whether the 1st or the 2nd voice in the voice pair matched the target voice by pressing "1" or "2" on the numerical keypad. The voices were presented at the same loudness for all participants. This was at level that was typical of a conversation you would hear in everyday life. Upon completion of the experiment, participants were fully debriefed.

Analyses

The results were analysed using mixed-group Analysis of Variance (ANOVA), one for the F0 manipulations and one for the speech rate manipulations. Owing to the high number of main effects and possible interactions, it was necessary to adjust the *p*-values from the main analysis to account for the familywise error rate. A Hochberg correction was therefore applied to the results of the main ANOVA (Hochberg, 1988). In addition, a Hochberg correction was applied to the simple main effects, which were conducted using pairwise *t*-tests. Furthermore, and for reasons of clarity, we present here only the significant findings of these analyses or where non-significant findings are directly relevant. Full ANOVA tables displaying the degrees of freedom (df), *F* ratios (*F*), effect sizes (generalised eta squared; η_G^2), and adjusted *p* values using the Hochberg correction (*p*) for all the main study variables and associated interactions are provided in Appendix E (*for F0*) and Appendix F (*for speech rate*).

Results

Fundamental frequency (F0)

Table 1 presents the mean percentage of errors made for each distractor type, listed separately for the three target conditions (high, moderate and low F0), the sex of the target voice, and listener sex.

The mean matching error scores for each listener were entered in a mixed ANOVA for the between subjects factor of listener sex (male or female) and the within subjects factors of sex of voice (male or female), target F0 (high, moderate or low), magnitude of distractor change (5%, 7%, or 10%) and direction of manipulation (increase or decrease in F0). This revealed a significant main effect of direction of manipulation, $F(1, 28) = 94.56$, $p < .03$, $\eta_G^2 = .07$, with significantly more errors being made when distractor voices were higher in F0 ($M = 21.20$, SD = 7.38) than when they were lower in F0 ($M = 10.51$, SD = 5.01).[3] There was also a significant main effect of magnitude of distractor change, $F(2, 56) = 50.75$, $p < .03$, $\eta_G^2 = .13$. Significantly more errors were made when distractor voices were manipulated by 5% ($M = 22.64$, SD = 7.63) compared to when they were manipulated by 7% ($M = 15.97$, SD = 7.31), $t(29) = 4.74$, $p < .001$, $d = 0.37$, and 10% ($M = 8.96$, SD = 5.66), $t(29) = 10.10$, $p < .001$, $d = 0.76$.[4] Significantly more errors were also made when distractor voices were manipulated by 7% ($M = 15.97$, SD = 7.31) compared to when they were manipulated by 10% ($M = 8.96$, SD = 5.66), $t(29) = 5.63$, $p < .001$, $d = 0.39$. No other main effects were significant or close to significance (adjusted $p > .93$).

In addition to the main effects, there was a significant interaction between target F0 and direction of manipulation, $F(2, 56) = 9.27$, $p < .05$, $\eta_G^2 = .03$.[5] As can be seen

Table 1. Mean percentage of errors made by distractor type (magnitude and direction of distractor change), target F0 (high, moderate (mod), or low), sex of target voice (collapsed across male and female target voices), and sex of listener (male or female).

	Male listener						Female listener					
	Male target voice			Female target voice			Male target voice			Female target voice		
	High	Mod	Low	High	Mod	Low	High	Mod	Low	High	Mod	Low
Distractor												
+10%	**6.67**	**6.67**	**16.67**	**3.33**	**6.67**	**6.67**	**8.33**	**13.33**	**21.67**	**6.67**	**15.00**	**13.33**
	14.84	*11.44*	*22.49*	*8.80*	*14.84*	*11.44*	*15.43*	*20.85*	*20.85*	*14.84*	*18.42*	*12.91*
+7%	**15.00**	**21.67**	**40.00**	**11.67**	**13.33**	**23.33**	**20.00**	**26.67**	**38.33**	**10.00**	**25.00**	**23.33**
	18.42	*20.85*	*28.03*	*18.58*	*12.91*	*22.09*	*23.53*	*22.09*	*24.76*	*18.42*	*29.88*	*14.84*
+5%	**20.00**	**25.00**	**38.33**	**18.33**	**23.33**	**38.33**	**55.00**	**30.00**	**30.00**	**30.00**	**21.67**	**40.00**
	21.55	*25.99*	*20.85*	*19.97*	*17.59*	*26.50*	*28.66*	*28.66*	*19.37*	*21.55*	*16.00*	*18.42*
−5%	**8.33**	**21.67**	**13.33**	**23.33**	**13.33**	**5.00**	**10.00**	**25.00**	**10.00**	**15.00**	**18.33**	**10.00**
	15.43	*18.58*	*18.58*	*22.09*	*16.00*	*10.35*	*15.81*	*23.15*	*15.81*	*12.68*	*17.59*	*18.42*
−7%	**1.67**	**10.00**	**13.33**	**3.33**	**11.67**	**8.33**	**10.00**	**18.33**	**10.00**	**6.67**	**11.67**	**10.00**
	6.46	*18.42*	*12.91*	*8.80*	*16.00*	*15.43*	*18.42*	*17.59*	*22.76*	*11.44*	*22.89*	*15.81*
−10%	**6.67**	**13.33**	**8.33**	**6.67**	**5.00**	**50.00**	**6.67**	**6.67**	**6.67**	**6.67**	**11.67**	**6.67**
	14.84	*18.58*	*18.09*	*14.84*	*10.35*	*14.02*	*11.44*	*14.84*	*14.84*	*14.84*	*16.00*	*11.44*

Note: Means are shown in bold. Standard deviations (SD) are shown in italics.

In Figure 1, the listeners selected higher F0 distractors more often than they selected lower F0 distractors. This effect was strongest for low F0 target voices, with more errors being made when target voices were paired with distractors higher in F0 ($M = 27.50$, SD $= 10.63$) than distractors lower in F0 ($M = 8.89$, SD $= 9.11$), $t(29) = 8.37$, $p < .001$, $d = 1.04$. A similar pattern of findings was apparent for high F0 target voices, with more errors being made when target voices were paired with distractors higher in F0 ($M = 17.08$, SD $= 12.09$) than distractors lower in F0 ($M = 8.75$, SD $= 7.48$), $t(29) = 3.73$, $p < .01$, $d = 0.46$. More errors were also made when moderate F0 target voices were paired with distractors higher in F0 ($M = 19.03$, SD $= 13.07$) than distractors lower in F0 ($M = 13.89$, SD $= 9.21$), $t(29) = 2.40$, $p < .05$, $d = 0.29$.

There was also a significant interaction between direction of manipulation and magnitude of distractor change, $F(2, 56) = 18.01$, $p < .03$, $\eta_G^2 = .04$.[6] Figure 2 shows that listeners selected distractor voices higher in F0 more often than they selected distractor voices lower in F0 when identifying target voices. This effect was strongest for distractor voices that sounded more similar in F0 to target voices. Specifically, listeners made more errors for distractor voices higher in F0 ($M = 30.83$, SD $= 11.24$) than distractor voices lower in F0 ($M = 14.44$, SD $= 7.24$) when distractor voices were manipulated by 5%, $t(29) = 8.05$, $p < .001$, $d = 0.91$. Listeners also made more errors for distractor voices higher in F0 ($M = 22.36$, SD $= 10.64$) than distractor voices lower in F0 ($M = 9.58$, SD $= 6.58$) when distractor voices were manipulated by 7%, $t(29) = 7.02$, $p < .001$, $d = 0.72$. A similar pattern of findings was also observed for distractor voices that sounded less similar in F0 to target voices (i.e. manipulated by 10%), with more errors being made

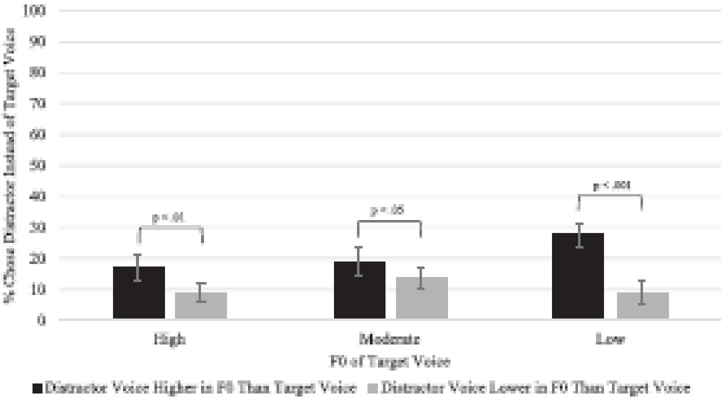


Figure 1. Mean percentage of errors made (i.e. chose distractor voice instead of target voice) for the three F0 target voice conditions. 95% confidence intervals are also shown.

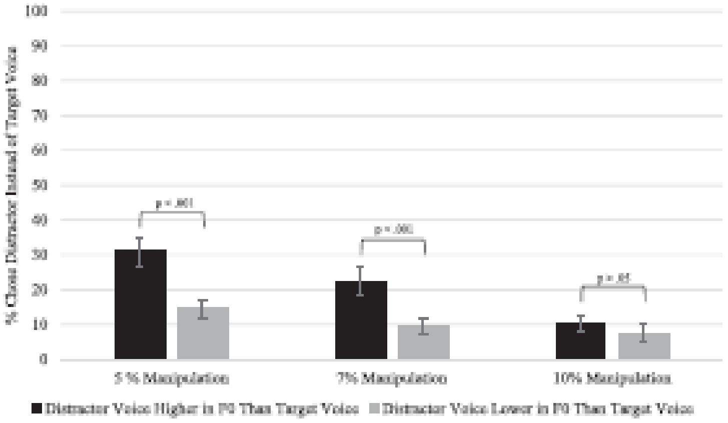


Figure 2. Mean percentage of errors made for F0 (i.e. chose distractor voice instead of target voice) for the 5%, 7%, and 10% distractor manipulations. 95% confidence intervals are also shown.

for distractor voices higher in F0 ($M = 10.42$, $SD = 6.17$) than distractor voices lower in F0 ($M = 7.50$, $SD = 7.37$), $t(29) = 2.13$, $p < .05$, $d = 0.16$.

No other interaction effects were significant or close to significance (adjusted $p > .31$).

Speech rate

For speech rate, the percentage of mean matching errors made for each distractor type, listed separately for each of the three target conditions (fast, moderate and slow speech rate), the sex of the target voice, and listener sex, are presented in Table 2.

The matching error scores for each listener were entered in a mixed ANOVA for the between subjects factor of listener sex (male or female) and the within subjects factors of sex of voice (male or female), target speech rate (fast, moderate or slow), magnitude of distractor change (10%, 12%, or 20%) and direction of manipulation (increase or decrease in rate). This revealed a significant main effect of direction of manipulation, $F(1, 28) = 12.55$, $p < .05$, $\eta_G^2 = .02$, with significantly more errors being made when the distractor voices were faster in speech rate ($M = 30.56$, $SD = 7.48$) than when they were slower in speech rate ($M = 25.05$, $SD = 8.06$). There was also a main effect of magnitude of distractor change, $F(1, 28) = 50.27$, $p < .05$, $\eta_G^2 = .10$. Significantly more errors were made when distractor voices were manipulated by 10% ($M = 33.69$, $SD = 8.35$) compared to when they were manipulated by 20% ($M = 18.75$, $SD = 7.95$), $t(29) = 9.90$, $p < .001$, $d = 0.65$. Significantly more errors were also made when distractor voices were manipulated by 12% ($M = 30.97$, $SD = 8.29$) compared to when they were manipulated by 20%

Table 2. Mean percentage of errors made by distractor type (magnitude and direction of distractor change), target speech rate (fast, moderate (mod), or slow), sex of target voice (collapsed across male and female target voices) and sex of listener (male or female).

	Male listener						Female listener					
	Male target voice			Female target voice			Male target voice			Female target voice		
	Fast	Mod	Slow	Fast	Mod	Slow	Fast	Mod	Slow	Fast	Mod	Slow
Distractor												
+20%	**20.00**	**11.67**	**35.00**	**30.17**	**10.00**	**33.33**	**15.00**	**18.33**	**30.00**	**25.00**	**18.33**	**30.00**
	21.55	*20.85*	*29.58*	*28.79*	*18.33*	*24.40*	*18.42*	*14.84*	*25.36*	*32.75*	*24.08*	*25.36*
+12%	**40.00**	**26.67**	**25.00**	**30.00**	**30.00**	**48.33**	**35.00**	**33.33**	**26.67**	**30.00**	**25.00**	**45.00**
	26.39	*22.09*	*16.37*	*27.06*	*28.66*	*22.09*	*18.42*	*28.16*	*17.59*	*22.08*	*18.90*	*19.37*
+10%	**33.33**	**43.33**	**41.67**	**30.17**	**30.00**	**51.67**	**38.33**	**35.00**	**40.00**	**30.17**	**26.67**	**45.00**
	26.16	*33.36*	*18.09*	*28.79*	*25.36*	*25.82*	*24.76*	*28.03*	*29.58*	*28.79*	*17.59*	*23.53*
−10%	**33.33**	**43.33**	**20.00**	**26.67**	**33.33**	**21.67**	**41.67**	**28.33**	**23.33**	**40.00**	**25.00**	**30.00**
	27.82	*29.07*	*21.55*	*22.09*	*26.16*	*22.89*	*34.93*	*28.50*	*25.82*	*22.78*	*18.90*	*19.37*
−12%	**38.33**	**28.33**	**16.67**	**35.00**	**20.00**	**36.67**	**35.00**	**23.33**	**23.33**	**38.33**	**20.00**	**35.00**
	20.85	*28.14*	*20.41*	*24.64*	*23.53*	*22.89*	*29.58*	*25.82*	*14.84*	*26.50*	*25.36*	*22.76*
−20%	**21.67**	**13.33**	**6.67**	**26.67**	**13.33**	**15.00**	**23.33**	**15.00**	**10.00**	**23.33**	**10.00**	**6.67**
	22.89	*12.91*	*14.84*	*24.03*	*20.85*	*18.42*	*17.59*	*22.76*	*20.70*	*19.97*	*15.67*	*11.44*

Note: Means are shown in bold. Standard deviations (SD) are shown in italics.

($M = 18.75$, SD = 7.95), $t(29) = 9.03$, $p < .001$, $d = 0.53$. However, there were no differences in errors made for distractor voices manipulated by 10% ($M = 33.69$, SD = 8.35) and 12% ($M = 30.97$, SD = 8.29), $t(29) = 1.52$, $p > .05$, $d = 0.12$. No other main effects were significant or close to significance (adjusted $p > .96$).

In addition to the main effects, there was also a significant interaction between target speech rate and direction of manipulation, $F(2, 56) = 15.12$, $p < .05$, $\eta_g^2 = .06$.[7] Figure 3 shows that for slow speech rate target voices, listeners selected distractors faster in rate ($M = 37.64$, SD = 11.13) more often than they selected distractors slower in rate ($M = 20.42$, SD = 10.68), $t(29) = 6.34$, $p < .001$, $d = 0.75$. However, there was no difference in the selection of distractors faster in rate ($M = 28.35$, SD = 14.86) and distractors slower in rate ($M = 31.94$, SD = 13.33) for fast speech rate target voices, $t(29) = -1.22$, $p > .05$, $d = 0.16$. Furthermore, there was no difference in the selection of distractors faster in rate ($M = 25.69$, SD = 13.97) and distractors slower in rate ($M = 22.78$, SD = 12.89) for moderate speech rate target voices, $t(29) = 1.20$, $p > .05$, $d = 0.13$.

No other interaction effects were significant or close to significance (adjusted $p > .43$).

Discussion

The current research investigated the impact of manipulations in F0 and speech rate on immediate target matching performance (selecting a voice from a pair to match a previously heard target voice) for a range of unfamiliar synthesised voices. We found that there was an increase in the selection of voices higher in F0 when high, moderate, and low F0 target voices were presented. For speech rate, there was an increase in the selection of voices faster in speech rate when slow speech rate target voices were presented. However, no such effect was detected for fast and moderate speech rate target voices. Therefore, in terms of our original hypotheses, there was no evidence for accentuation effects for voice memory. Furthermore, for both the F0 and speech rate conditions, more errors were made identifying target voices when paired with distractor voices manipulated by a smaller magnitude (i.e. 5% for F0, and 10% for speech rate) compared to those manipulated by a greater magnitude (i.e. 10% for F0, and 20% for speech rate). This is perhaps unsurprising given that the results from the pilot study suggest that voices manipulated by a smaller magnitude are harder to distinguish between, and sound more similar, to original voices than voices manipulated by a greater magnitude. Thus, more errors are likely to be made identifying target voices when paired with distractor voices manipulated by a smaller magnitude because any differences between the voices are more difficult to detect. There was no effect of either sex of voice or listener sex on errors made identifying target voices.

Fundamental frequency (F0)

The results presented here do offer some support to those identified by Mullenix et al. (2010) in that errors in memory are likely to occur for voice F0. However, the finding of an increase in the selection of voices higher in F0 is difficult to explain using the accentuation effect alone. We believe that this outcome is not an

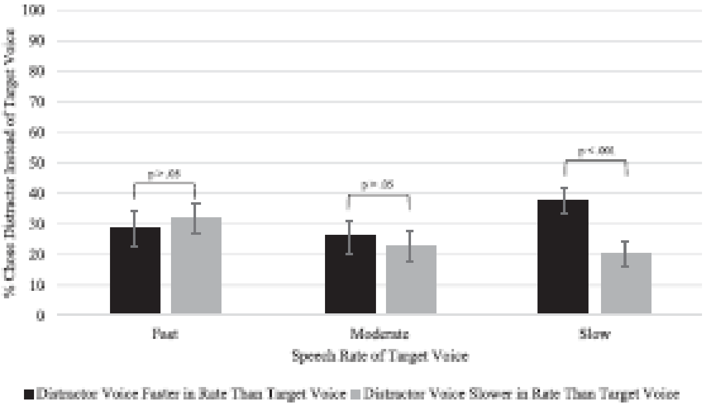


Figure 3. Mean percentage of errors made (i.e. chose distractor voice instead of target voice) for the three speech rate target voice conditions. 95% confidence intervals are also shown.

anomaly in our data set given that the findings are reasonably consistent across all target voices. They are also unlikely to be the result of order effects because we counterbalanced the voices that were presented to listeners in the voice pair. Given that synthesised voices were used for experimentation, we do acknowledge that some of the acoustic properties of the stimuli could explain the observed pattern of findings. However, we believe this is unlikely given that the voices used were rated as sounding natural, formant frequencies changed freely, formant transitions were smooth, and there were no intonational irregularities or prosodic mismatches across words. This alleviates concerns that something uncontrolled and artificial about the stimuli were driving the findings. Rather, we propose that the extensions and modifications made to the study procedure may explain the difference in results. First, we kept the target and distractor voices within a F0 range that is typical in the population (i.e. between 80 and 180 Hz, and for an adult female this will be between 165 and 255 Hz (Titze, 1994). In contrast, the manipulations made by Mullenix et al. (2010) fell considerably outside of this range. Second, we used a set of four synthesised voices, whereas Mullenix et al. (2010) used only a single voice. Therefore, it is quite possible that the findings identified by Mullenix et al. (2010) were due to the peculiarity of the stimuli (i.e. an unusually high or low F0) used in the experiment. Using a more representative and generalisable set of voices, as in the present study (i.e. a slightly larger set of synthesised voices, with manipulations in F0 and speech rate kept within a range that is typical in the population for English speakers), the accentuation bias is no longer found. The data reported here suggest little or no accentuation bias for the memory of voice F0.

Why then do listeners make more errors recognising voice F0 when paired with distractor voices higher in F0 compared to when they are paired with distractor voices lower in F0? It is quite possible that listeners had difficulty discriminating between the frequencies of some of the voice pairs in the experiment. Indeed, research has identified that it is more difficult to discriminate between voices of higher frequencies compared to voices of lower frequencies (Moore, 1995). In the present study, listeners may have made fewer errors identifying target voices when paired with distractor voices lower in F0 because they were more efficient at detecting the changes in frequency than when distractor voices were higher in frequency. This interpretation would account for why there was no effect of listener sex on errors made identifying target voices, because there is no reason to believe that the perceptual capabilities of the listener would differ substantially between male and female listeners. It would also explain why there was no difference in errors made for male and female target voices. Although female voices are higher in F0 than male voices, the findings are based upon a listener's ability to detect any *differences* in the frequencies of the voices in the voice pair, and this is independent of the frequency of the target voice itself.

It is also likely that listeners made more errors identifying target voices when paired with distractor voices higher in F0 compared to when they are paired with distractor voices lower in F0 because they resemble voices that are typically heard in the general population. Inflection refers to the frequency patterns in a person's speech, where the voice rises and falls, either upwards or downwards in frequency (Fairbanks, 1940). Research has shown that all types of inflections are greater in upward inflection than they are in downward inflection (e.g. Fairbanks & Pronovost, 1939). Furthermore, researchers have shown that when people are asked to choose a method of disguise, they are more likely to raise the frequency of their voice rather than lowering it (e.g. Masthoff, 1996; Mathur, Choudhary, & Vyas, 2013). Such evidence suggests that people are more likely to increase, rather than decrease, the frequency of their voice when they speak. Thus, the listeners in the present study may be selecting distractor voices higher in F0 more often than distractor voices lower in F0 because they are more familiar with these types of utterances and it sounds like a more plausible version of the target voice (i.e. an inflected version of the target voice).

The finding that listeners were more likely to select distractor voices higher in F0 compared to distractor voices lower in F0 was particularly prevalent for the low F0 target voice condition. This bias may have arisen because voices higher in F0 are perceived as less threatening than voices lower in F0. Research has shown that both male and female voices lowered in F0 are perceived as more dominant than the same voices raised in F0 (Borkowska & Pawlowski, 2011; Fraccaro et al., 2013; Jones, Feinberg, DeBruine, Little, & Vukovic, 2010; Puts, Gaulin, & Verdonili, 2006). Furthermore, evidence tends to suggest that people will often exhibit avoidance type behaviour when exposed to aversive stimuli (Corr, 2013). Assuming that voices lower in F0 would be rated as more dominant and threatening than the voices higher in F0 in the present study, listeners may have selected the higher voice of the pair because it sounded less dominant and less threatening. This interpretation would explain why an increase in the selection of higher F0 distractors was particularly prevalent for the low F0 target voice condition; because the voices were decreased in F0 sufficiently for the higher F0 voices in the pair to be perceived as less threatening to the listener. It would also account for why there was no effect of either sex of voice or listener sex; perceptions of dominance have been found to be equivalent for both male and female voices and male and female listeners (Jones et al., 2010). Further work would be required to confirm or disconfirm this explanation to our finding. Another possibility that also deserves equal consideration is that English voices lower in F0 for both males and females tend to co-occur with covariations in voice quality (e.g. Abberton, Howard, & Fourcin, 1989). A bias towards selecting the higher F0 distractor voices could reflect the unnaturalness of the voices lowered in F0 without a concomitant change in voice quality. Whilst the voices were rated as sounding natural, this issue might still remain even if naturally sounding voices were modified to have a lower F0.

Finally, as pointed out by one reviewer, for which we are grateful, it is worth noting that the naturalness ratings for the voices with higher F0 manipulations tended to yield slightly higher naturalness rating scores than those with lower manipulations (refer to Appendix C for further details). One possible interpretation of this is that the listeners preferred the more natural sounding voices (i.e. the higher F0 manipulations) and were thus, more likely to select them. Unfortunately, because the naturalness ratings came from a different population to those in the 2AFC tasks reported here,it was not appropriate to formally test this possibility. Our tuition, given that the voices were generally perceived to be natural sounding across the board, the differences observed between the voices being relatively small, are unlikely to have impacted upon the matching tasks. Thus, whilst we accept it is a possibility that naturalness may have an effect, we are unable to resolve the question here.

Speech rate

For speech rate, listeners selected voices faster in speech rate when slow speech rate target voices were presented. Thus, the findings presented here cannot be explained using the accentuation effect. Given this, it is possible that the findings could be accounted for by the listener's level of familiarity of the voice heard. In natural speech, a person speaking more slowly is likely to be more hesitant, making more silent pauses or filled pauses (e.g. *um*, *er*). In the present study, a decreasing speech rate did affect the rate of continuous production but did not lead to increased pauses of any kind. It is therefore unlikely that the speech samples used were an entirely natural rendition of slower speech, at least of a type that listeners most typically hear. It is possible that at the lower margins of the speech rate manipulated samples (i.e. the slowest samples), but not elsewhere, the participants may have selected a faster voice in the pair because it sounded more realistic.

Faster speaking voices might also sound more favourable when compared with slower speaking voices in the slow speech rate pairings. Indeed, research suggests that speech rates can influence a listener's perceptions of a speaker's personality and social skills. For example, faster speaking styles have been shown to be rated more favourably (Stewart & Ryan, 1982), and viewed as more competent and socially attractive than voices spoken at a slower rate (Street, Brady, & Putman, 1983). Slower speaking styles have also been identified as sounding weaker, less truthful, and less empathetic than voices spoken at a faster rate (Apple, Streeter, & Krauss, 1979). It is possible that listeners were more likely to select a faster voice in the pair because they preferred the sound of the voice. However, such selections may have been made only for the slow speech rate condition because these voices were slowed sufficiently for the faster rate voices in the pair to be rated more favourably, and thus selected by the listener. The above explanations would also account for why there was no effect of either sex of voice or listener sex on errors made identifying a target voice, as there is no reason to suggest that the level of familiarity or preference for faster voices would differ between male and female voices, or for male and female listeners.

Concluding comments

The results from the present study suggest that, at least for synthesised voices, listeners are susceptible to distortions in memory for certain properties of the voice more so than others. However, the accentuation bias does not account for our findings here. Therefore, it is doubtful that listeners rely solely on the categorical information self-generated about the voice at the time of encoding to aid in recognition of the voice at a later stage. The present study has thus contributed to our understanding of the mechanisms important for accurate voice recognition and such work may prove as a useful conceptual tool in determining the properties of voice that are more or less affected by intra-individual variation. Future work in this field should focus on framing their research with a more applied perspective in mind. For example, it would be particularly valuable to establish whether the results from the present study would also extend to real, rather than synthesised voices. Future work could also be undertaken to determine the impact of longer retention intervals on errors made identifying voices. This is especially interesting given that in a real world criminal situation there is uncertainty over the time period between hearing a voice and being asked to identify the voice at a later date. Such work would undoubtedly advance on the research currently being carried out in this domain and further our understanding of the impact of manipulations of certain characteristics of our voice, whether it be through unintentional or deliberate means.

Notes

1. Calculations using syllables rather than words are often considered as being a more accurate and reliable estimate of the rate of speech (Dlugan, 2012). This is because calculations using words are dependent upon the length of the words spoken in the spoken sentence, and not all words in the English language are equal. It was therefore decided upon to use syllables per second (syll/sec) for all calculations of speech rate.
2. It should be noted that formant values changed freely as a result of manipulations in F0. Changes in formants would occur naturally in real voices when changes in F0 are made. This helped to retain the naturalness of the voices used by limiting any irregularities that might by introduced in the voices via the use of synthesised speech (refer to Appendix A for further details about the formant values of the voices).
3. Generalised eta-squared statistics (η_G^2) are reported here in order to facilitate comparison between studies with different designs. Generalised eta-square describes the proportion of sample variance accounted for by an effect in an independent design with no manipulated factors (Olejnik & Algina, 2003).
4. Cohen's *d* values were determined by calculating the mean difference between the two groups, and then dividing the result by the overall pooled standard deviation from all conditions (for F0 = 17.95, for speech rate = 22.98).
5. The tests of simple main effects that follow are again adjusted using the Hochberg correction. Note that the Hochberg correction is not conditional on a significant *F* ratio in order to protect the Type 1 error. We corrected for all six possible simple main effects. However, for reasons of brevity we report here only the three simple main effects that are of direct interest.
6. We corrected for all nine possible simple main effects. However, for reasons of brevity we report here only the three simple main effects that are of direct interest.
7. We corrected for all nine possible simple main effects. However, for reasons of brevity we report here only the three simple main effects that are of direct interest.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Vice Chancellor Bursary - Nottingham Trent University.

References

Abberton, E. R. M., Howard, D. M., & Fourcin, A. J. (1989). Laryngographic assessment of normal voice: A tutorial. *Clinical Linguistics and Phonetics*, *3*, 281–296. doi:10.3109/02699208908985291

Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, *37*, 715–727. doi:10.1037/0022-3514.37.5.715

Arnfield, S., Roach, P., Setter, J., Greasley, P., & Horton, D. (1995). *Emotional stress and speech tempo variation*. Processings of ESCA-NATO Tutorial and Research Workshop on Speech Under Stress, Lisbon, 13–15.

Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, *82*, 55–59. doi:10.1016/j.anbehav.2011.03.024

Brosch, T., Pourtois, G., & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion*, *24*, 377–400. doi:10.1080/02699930902975754

Brown, A. (2014). *Pronunciation and phonetics: A practical guide for English language teachers*. New York: Routledge. doi:10.1075/jslp.1.2.07mac

Corr, P. J. (2013). Approach and avoidance behaviour: Multiple systems and their interactions. *Emotion Review*, *5*(3), 285–290. doi:10.1177/1754073913477507

Corneille, O., Huart, J., Becquart, E., & Brédart, S. (2004). When memory shifts toward more typical category exemplars: Accentuation effects in the recollection of ethnically ambiguous faces. *Journal of Personality and Social Psychology*, *86*, 236–250. doi:10.1037/0022-3514.86.2.236

Dlugan, A. (2012). *What is the average speaking rate?* Retrieved from http://sixminutes.dlugan.com/speaking-rate/#author

Eiser, J. R. (1971). Enhancement of contrast in the absolute judgment of attitude statements.. *Journal of Personality and Social Psychology*, *17*, 1–10. doi:10.1037/h0030455

Endres, W., Bambach, W., & Flosser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. *The Journal of The Acoustical Society of America*, *49*, 1842–1848. doi:10.1121/1.1912589

Fairbanks, G. (1940). Recent experimental investigations of vocal pitch in speech. *The Journal of the Acoustical Society of America*, *11*(4), 457–466. doi:10.1121/1.1916060

Fairbanks, G., & Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion. *Communications Monographs*, *6*(1), 87–104. doi:10.1080/03637753909374863

Fiske, S. T., Gilbert, D. T., & Lindzey, G. (2010). *Handbook of social psychology: Volume one*. NJ: John Wiley & Sons. doi:10.1002/9780470561119

Fraccaro, P. J., O'Connor, J. J., Re, D. E., Jones, B. C., DeBruine, L. M., & Feinberg, D. R. (2013). Faking it: Deliberately altered voice pitch and vocal attractiveness. *Animal Behaviour*, *85*, 127–136. doi:10.1016/j.anbehav.2012.10.016

Halberstadt, J. B., & Niedenthal, P. M. (2001). Effects of emotion concepts on perceptual memory for emotional expressions. *Journal of Personality and Social Psychology*, *81*, 587–598. doi:10.1037/0022-3514.81.4.587

Haslam, S. A., & Turner, J. C. (1992). Context-dependent variation in social stereotyping 2: The relationship between frame of reference, self-categorization and accentuation.

European Journal of Social Psychology, *22*, 251–277. doi:10.1002/ejsp.2420220305

Herlitz, A., Nilsson, L. G., & Backman, L. (1997). Gender differences in episodic memory. *Memory and Cognition*, *25*, 801–811. doi:10.3758/bf03211324

Hillar, K.F., & Kemp, R.I. (2008). Barak Obama or Barry Dunham? The appearance of multiracial faces is affected by the names assigned to them. *Perception*, *37*, 1605–1608. doi:10.1068/p6255

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*, 800–802. doi:10.1093/biomet/75.4.800

Hogg, M., & Vaughan, G. (2010). *Essentials of social psychology*. Harlow: Pearson Education Limited.

Hollien, H. (1990). *The Acoustics of Crime: The New Science of Forensic Phonetics*. New York: Springer.

Huart, J., Corneille, O., & Becquart, E. (2005). Face-based categorization, context-based categorization, and distortions in the recollection of gender ambiguous faces. *Journal of Experimental Social Psychology*, *41*, 598–608. doi:10.1016/j.jesp.2004.10.007

Jones, B. C., Feinberg, D., DeBruine, L. M., Little, A. C., & Vukovic, J. (2010). A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour*, *79*, 57–62. doi:10.1016/j.anbehav.2009.10.003

Jreige, C., Patel, R., & Bunnell, H. T. (2009, October). VocaliD: Personalizing text-to-speech synthesis for individuals with severe speech impairment. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility* (pp. 259–260). ACM. doi:10.1145/1639642.1639704

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology*, *67*, 596–610. doi:10.1037/0022-3514.67.4.596

Krueger, J., & Rothbart, M. (1990). Contrast and accentuation effects in category learning. *Journal of Personality and Social Psychology*, *59*, 651–663. doi:10.1037/0022-3514.59.4.651

Levin, D. T., & Banaji, M. R. (2006). Distortions in the perceived lightness of faces: The role of race categories. *Journal of Experimental Psychology: General*, *135*, 501–512. doi:10.1037/0096-3445.135.4.501

Lewin, C., Wolgers, G., & Herlitz, A. (2001). Sex differences favouring women in verbal but not visuospatial episodic memory. *Neuropsychology*, *15*, 165–173. doi:10.1037/0894-4105.15.2.165

MacLin, O. H., & Malpass, R. S. (2001). Racial categorization of faces: The ambiguous race face effect. *Psychology, Public Policy, and Law*, *7*, 98–118. doi:10.1037/1076-8971.7.1.98

Masthoff, H. (1996). A report on a voice disguise experiment. *Forensic Linguistics*, *3*, 160–167. doi:10.1558/ijsll.v3i1.160

Mathur, S., Choudhary, S. K., & Vyas, J. M. (2013). Speaker recognition system and its forensic implications. *Open Access Scientific Reports*, *2*(4), 723. doi:10.4172/scientificreports.723

McGarty, C., & Penny, R. E. C. (1988). Categorization, accentuation and social judgement. *British Journal of Social Psychology*, *27*, 147–157. doi:10.1111/j.2044-8309.1988.tb00813.x

McGarty, C., & Turner, J.C. (1992). The effects of categorization on social judgement. *British Journal of Social Psychology*, *31*, 253–268. doi:10.1111/j.2044-8309.1992.tb00971.x

McGivern, R. F., Huston, J. P., Byrd, D., King, T., Siegle, G. J., & Reilly, J. (1997). Sex differences in visual recognition memory: Support for sex-related difference in attention in adults and children. *Brain and Cognition*, *34*, 323–336. doi:10.1006/brcg.1997.0872

Moore, B. C. J. (1995). *Hearing*. San Diego, CA: Academic Press.

Mullenix, J. W., Stern, S. E., Grounds, B., Kalas, R., Flaherty, M., Kowalok, S., ... Tessmer, B. (2010). Earwitness memory: Distortions for voice pitch and speaking rate. *Applied Cognitive Psychology*, *24*, 513–526. doi:10.1002/acp.1566

Nolan, F. (2005). Forensic speaker identification and the phonetic description of voice quality. In W. J. Hardcastle & J. Mackenzie Beck (Eds.), *A figure of speech* (pp. 385–411). New York: Routledge.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*, 434–447. doi:10.1037/1082-989X.8.4.434

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13. doi:10.1016/j.jneumeth.2006.11.017

Puts, D. A., Gaulin, S. J. C., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, *27*, 283–296. doi:10.1016/j.evolhumbehav.2005.11.003

Queller, S., Schell, T., & Mason, W. (2006). A novel view of between-categories contrast and within-category assimilation. *Journal of Personality and Social Psychology*, *91*, 406–422. doi:10.1037/0022-3514.91.3.406

Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguise upon speaker identification by listening. *The Journal of The Acoustical Society of America*, *66*, 1023–1028. doi:10.1121/1.383321

Reich, A. R., Moll, K. L., & Curtis, J. F. (1976). Effects of selected vocal disguises upon spectrographic speaker identification. *The Journal of The Acoustical Society of America*, *60*, 919–925. doi:10.1121/1.2003461

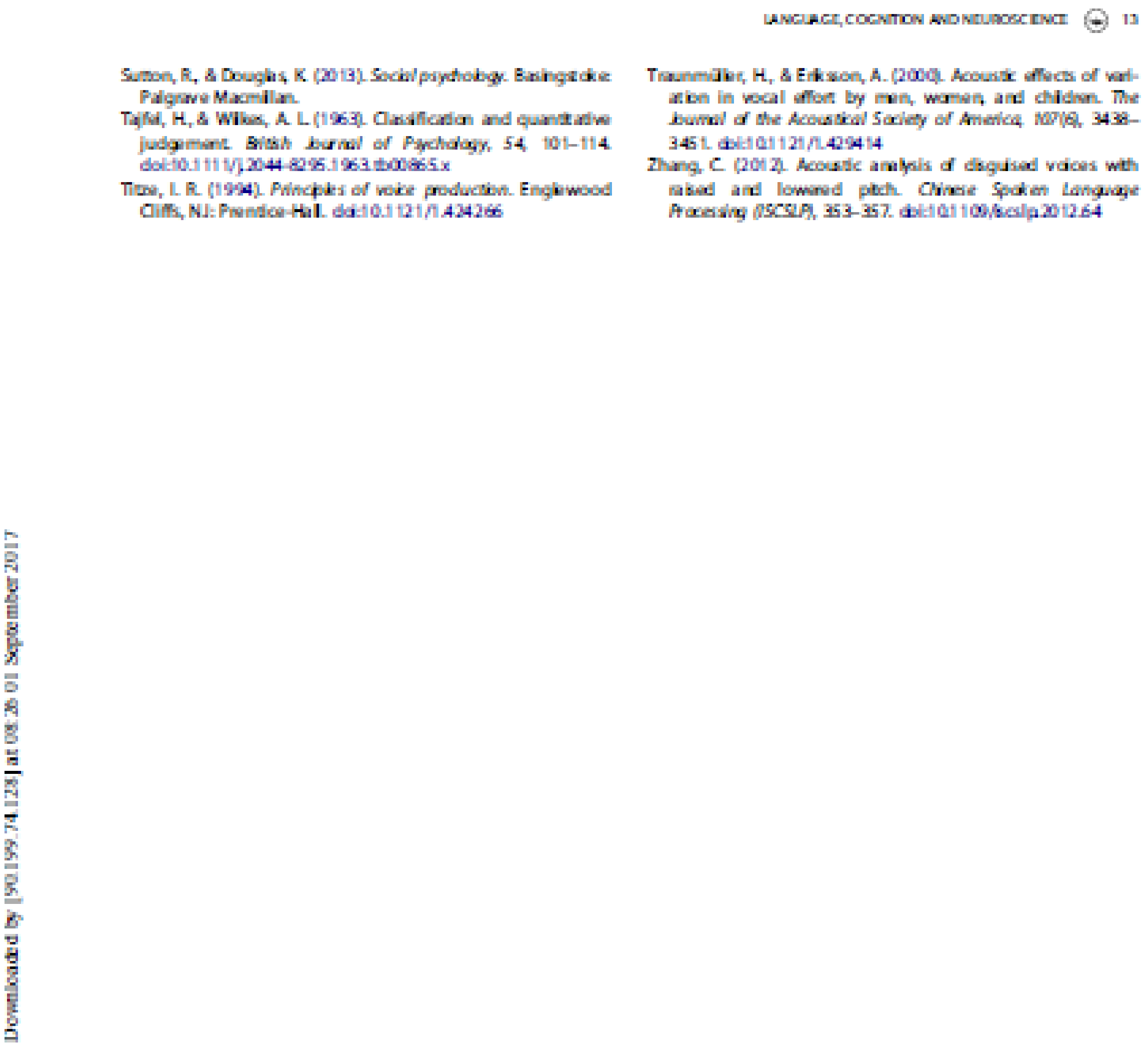
Roebuck, R., & Wilding, J. (1993). Effects of vowel variety and sample length on identification of a speaker in a line-up. *Applied Cognitive Psychology*, *7*, 475–481. doi:10.1002/acp.2350070603

Saslove, H., & Yarmey, A.D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology*, *65*, 111–116. doi:10.1037/0021-9010.65.1.111

Stern, S. E., Mullennix, J. W., Corneille, O., & Huart, J. (2007). Distortions in the memory of the pitch of speech. *Experimental Psychology*, *54*, 148–160. doi:10.1027/1618-3169.54.2.148

Stewart, M. A., & Ryan, E. B. (1982). Attitudes toward younger and older adult speakers: Effects of varying speech rates. *Journal of Language and Social Psychology*, *1*, 91–109. doi:10.1177/0261927x8200100201

Street, R. L., Brady, R. M., & Putman, W. B. (1983). The influence of speech rate stereotypes and rate similarity or listeners' evaluations of speakers. *Journal of Language and Social Psychology*, *2*, 37–56. doi:10.1177/0261927x8300200103



Sutton, R., & Douglas, K. (2013). *Social psychology*. Basingstoke: Palgrave Macmillan.

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British Journal of Psychology*, *54*, 101–114. doi:10.1111/j.2044-8295.1963.tb00865.x

Titze, I. R. (1994). *Principles of voice production*. Englewood Cliffs, NJ: Prentice-Hall. doi:10.1121/1.424266

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*(6), 3438–3451. doi:10.1121/1.429414

Zhang, C. (2012). Acoustic analysis of disguised voices with raised and lowered pitch. *Chinese Spoken Language Processing (ISCSLP)*, 353–357. doi:10.1109/iscslp.2012.64