FAMILIAR TALKERS IN SPEECH PERCEPTION

KNOWING THE VOICE: FAMILIAR TALKERS IN SPEECH PERCEPTION

By CHRISTOPHER ARUFFO, B.S., M.B.A., M.F.A., M.Sc.

A thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the degree of Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY McMaster University

Psychology, Neuroscience & Behaviour Hamilton, Ontario

TITLE: Knowing the voice: Familiar talkers in speech perception

AUTHOR: Christopher Aruffo, B.S., M.B.A., M.F.A., M.Sc.

SUPERVISOR: Dr. David I. Shore

NUMBER OF PAGES: xiii, 192

Abstract

Millions of different people talk to each other, and no two people sound exactly the same. Yet, whomever we are listening to, we expect to easily understand what he or she has to say. Somehow, we adjust to each new talker's voice and hear the "same" speech sounds. Until recently, differences between voices were viewed as a perceptual problem interfering with speech perception. Recent developments, however, have shown that familiar voices can facilitate speech. Speech-perception models can no longer dismiss talkers' voices merely as carriers for speech, and models currently struggle to understand the relation between vocal identity and the content of speech. The present thesis contributed to this discussion by examining familiar talkers, whose identities have been encoded into listeners' memories. Chapter 2 studied familiar faces' and voices' contribution to audiovisual speech processing, and found that different listeners may focus more strongly on learning either a familiar talker's face or voice, but will recall what they have learned in response to that talker's voice, not face. Chapter 3 examined self-speech, and discovered that we do receive a familiar-talker speech-processing advantage from hearing our own recorded voice, but only so far as we can identify self-voice; when voices are obscured by noise, we receive an equivalent advantage for all voices of our own sex. Chapter 4 confirmed a relation between speech familiarity and accurate talker identification. Taken together, the data presented in this thesis support a model of speech perception in which listeners encode talkers' identities inclusive of both idiosyncratic speech production and vocal qualities and, when processing speech, recall

as many of a talker's identifying characteristics as can be usefully applied to an incoming speech signal. These findings contribute to our understanding of how we utilize talker identity in perceiving speech.

This dissertation is dedicated to my wonderful wife, Caitlin,

who has remained patient and supportive throughout.

Acknowledgements

I am grateful to David I. Shore, my advisor, who saw my potential at the start and, with his unflagging confidence and support, has helped me grow and learn as a scientist and an educator throughout this experience.

My committee members, Karin Humphreys, Joe Kim, and David Feinberg, I thank for their attention to and interest in my work, as well as their ongoing advisement, criticism, and encouragement.

Thanks to Bruce Milliken for always having sage words to offer, and to Lee Brooks for offering crucial perspective to my investigations.

I also thank David Earn for opening the door to this adventure, and Rob Goldstone for laying its groundwork.

For their contributions to my work and as part of the lab, thanks to Michelle Cadieux, Swapna Krishnamoorthy, Bruce Sheng, Katherine Jongsma, and Kyla Baird.

Thanks to my father, who was always there to help me express my thoughts effectively.

And thanks particularly to my mother, who has for decades been patiently but consistently supportive as she has waited for me to find this career.

Table of Contents

List of Figures

List of Tables

Declaration of Academic Achievement

Chapter 1: Introduction

Author: Christopher Aruffo

Chapter 2: Is a familiar face or a familiar voice more important to audiovisual speech?

Authors: Christopher Aruffo & David I. Shore

Publication: Submitted

Comments: This manuscript was conceived by CA and DIS. Data were collected and organized by CA and Katherine Jongsma. The manuscript was written by CA under the supervision of DIS.

Chapter 3: What'd I say? Word-recognition accuracy for self-speech

Authors: Christopher Aruffo & David I. Shore

Publication: Not submitted

Comments: This manuscript was conceived by CA and DIS. Data were collected and organized by CA. The manuscript was written by CA under the supervision of DIS.

Chapter 4: Identifying Familiar Speech

Authors: Christopher Aruffo & David I. Shore

Publication: Not submitted

Comments: This manuscript was conceived by CA and DIS. Data were collected and organized by CA. The manuscript was written by CA under the supervision of DIS.

Chapter 5: Discussion

Author: Christopher Aruffo

Chapter 1: Introduction

It might seem intuitively obvious that a familiar talker would be easier to understand. One might call to mind a daughter, at a restaurant, easily "translating" her father's thick foreign accent to a baffled waiter. Because the daughter has had many years of practice listening to her father, it seems plausible that she finds it easier to process her father's unusual speech patterns. But to do this, she would need to have learned, and stored in memory, the idiosyncratic auditory characteristics of her father's speech production. Only within the past few decades have we formally acknowledged that this could actually happen. Previously, identifying a voice and recognizing its speech were considered separate processes; it was supposed that listeners understood speech by extracting a voice's "speech" characteristics and processing those separately from the voice's "identity" characteristics. As evidence began to build showing that remembering a voice could facilitate later speech processing, it became increasingly clear that the identifying characteristics of a voice could not be fully separated from those that represented its speech. Models of speech perception began to address the role and function of talker identity in speech processing. Familiar talkers were of particular interest, because a familiar voice is one which a listener has encoded into memory. The differences between how listeners respond to familiar talkers versus unfamiliar talkers can help us understand how we encode talker identity with respect to speech. To examine how we remember and recall talker identity in speech processing, this thesis presented three investigations of familiar talkers.

1.1 Development of speech-processing models

Until the mid-1900s, speech sounds were assumed to be the natural result of mouth movements, or speech *articulation* (Joos, 1948). The process of recognizing speech was, therefore, thought to be a matter of recognizing which articulations had been produced. Accordingly, the "sounds" of a language were described as physical positions enacted by the articulators (e.g., tongue, lips, glottis). However, as technology became better able to analyze sound waves, it was discovered that the same articulatory actions, performed by different talkers, produced different acoustic patterns. Moreover, not all of the vibratory-frequency components produced by a voice were relevant to speech.

Indexical information, or talker-specific vocal components not directly implicated in speech production, was hypothesized to convey talker identity (McGehee, 1944). Indexical information comprises talker-specific features and habits, such as vocal-tract size and shape (Peterson & Barney, 1952; Schwartz, 1968; Schwartz & Rine, 1968), talker age (Shipp, Qi, Huntley, & Hollien, 1992; Harnsberger et al., 2008), glottal activity (Carr & Trill, 1964; Monsen & Engebretson, 1977), and language accent (Bradlow & Bent, 2008). Each indexical feature affects the acoustic realization of speech, with the most prominent talker-specific cues related to the shape of a talker's vocal apparatus (Baumann & Belin, 2010). Because indexical information was believed to be related to identity alone, the process of speech perception was thought to be the "extraction" of speech components from a vocal stream (cf. Studdert-Kennedy, 1980). That is, indexical information was considered a mere carrier of speech information; to understand a talker, listeners streamed language away from indexical information to reveal its speech targets.

Researchers therefore turned their attention toward how listeners solved the perceptual problem of extracting speech content from its indexical guises (e.g., Gerstman, 1968; Shearme & Holmes, 1959).

Talker-specific vocal features, it was supposed, functioned as a frame of reference against which a talker's speech patterns could be *normalized*. To normalize speech, listeners stripped away talker-specific vocal elements to extract standard abstract speech forms (cf. Johnson, 2005). Models of speech perception addressed indexical features in terms of their contribution to normalization (e.g., McClelland & Elman, 1986; Norris, 1994); for example, listeners recognized vowel sounds by calculating the relationship between their vibratory frequencies and the "fundamental" frequency produced by a talker's vocal folds (Hillenbrand, Getty, Clark, & Wheeler, 1995). According to the normalization model, talker identity was essentially noise obscuring the speech signal; speech was extracted and understood by identifying, segregating, and discarding all information related to talker identity (Halle, 1985).

1.2 Studies of Vocal Identity

Remembering a familiar voice was not considered a linguistic matter, but a forensic one. The study of vocal identity focused primarily on issues in "earwitness" testimony (McGehee, 1937). Research questions were designed to determine how effectively and reliably a voice could be identified, primarily testing the parameters within which memory for voices could be expected to be accurate (for a review, see Clifford, 1980). It was noted that voices were easier to identify when samples presented

more speech content (e.g., Pollack, Pickett, & Sumby, 1954; Bricker & Pruzansky, 1966), suggesting a link between identity and speech, but interest remained principally vested in how well an individual listener could remember a particular voice. Talker identification gained more practical application with advancing technology, such as electronic devices that could be activated, manipulated, or unlocked by voice (Doddington, 1985), but again it was the utility and reliability of voice identification that remained the most important research question, with specific investigations attempting to determine which acoustic qualities of a voice were the most definitive.

1.3 Effects of Familiar Voices

Previously heard voices conveyed speech-processing benefits, but these effects were interpreted as acoustic memory rather than an effect of talker identity. That is, listeners had remembered the details of discrete acoustic events, and same-voice speech facilitation could be attributed to the "representational persistence" of having heard the words spoken in that voice before (Cole, Coltheart, & Allard, 1974; Craik & Kirsner, 1974). However, experiments in auditory priming demonstrated a facilitation effect for previously-heard voices that was independent of semantic judgment (Church & Schacter, 1994; Schacter & Church, 1992), indicating that talker-specific information could be retained and applied to speech processing. Moreover, when listeners were trained to become familiar with a set of voices, those listeners became better able to recognize novel words from those same talkers (Nygaard, Sommers, & Pisoni, 1994), demonstrating that talker-specific speech information was not merely stored as the

acoustic details of a prior event, but was encoded as specific to that talker's speech production.

It is now clear that speech is processed more quickly, and recalled more accurately, when presented in a familiar voice, and that this effect may be attributed to the encoding of talker-specific information for use in speech processing (Bradlow, Nygaard, & Pisoni, 1999; Goh, 2005; Kraljic & Samuel, 2006; Luce & Lyons, 1998; Markham & Hazan, 2004; Palmeri, Goldinger, & Pisoni, 1993), especially because the more familiar the voice, the more effectively it facilitates speech processing (Magnuson, Yamada, & Nusbaum, 1995). This evidence has challenged the traditional view of speech normalization, in which talker-identifying features are merely a source of noise to be stripped away and nullified. New models now struggle to understand how talker identity is represented in speech processing. Studying listeners' responses to speech spoken by familiar voices, presented in different forms and contexts, may further that understanding.

1.4 Familiar Audiovisual Speech

Familiar speech can be auditory or visual. When we listen to people, in person, we see their faces and hear their voices, and we integrate their faces with their voices to perceive *audiovisual speech*. But the facial input is a separate sensory channel from the verbal input, and we can become separately familiar with either. We can become familiar with "visual speech" produced by a silently-talking face, and then lip-read that face more accurately (Lander & Davies, 2008), or we can become familiar with auditory speech

produced by an unseen talker, and then recognize that voice's speech more accurately (Nygaard, Sommers, & Pisoni, 1994). Yet being familiar with visual speech conveys an advantage to subsequent presentation of auditory-only speech (Rosenblum, Miller, & Sanchez, 2007); and being familiar with auditory speech can convey a lip-reading advantage (Sanchez, Dias, & Rosenblum, 2013). Familiarity with either sensory channel, therefore, crosses over into the other channel; and, because it is generally easier to recognize and recall a face than a voice (Hanley & Damjanovic, 2009), it may be that a familiar face has a greater influence on audiovisual speech than a familiar voice. This possibility raises the question of whether a face or voice makes a greater contribution to audiovisual speech processing.

The relative contribution of a face or a voice to audiovisual speech can be measured by using an audiovisual speech illusion known as the *McGurk effect*. The integration of faces and voices is accomplished by the principle of *optimal integration* (cf. Ernst & Bülthoff, 2004). That is, upon receiving the auditory and visual inputs of a speech signal, those inputs are "weighted" according to their apparent reliability. For example, a voice obscured by white noise might be assigned a lesser weight, or a brightly-lit face might be assigned a greater weight. We then integrate the weighted inputs into a final percept, each input contributing with mathematical predictability according to its assigned weight (Massaro, 2004). When faces and voices are equally reliable, each provides an equivalent contribution. This can produce a curious result when their contents conflict. When a voice speaking one sound, such as /aba/, is dubbed onto a face speaking another sound, such as /aga/, the two channels blend to produce a

third sound not actually present, such as /ada/. However, if one channel is more reliable than the other, the channels will be differently weighted, and the blend does not occur. Instead, a listener hears only the signal from the more heavily-weighted channel. Therefore, if a familiar face or voice facilitates speech processing, then introducing a familiar face or voice into either channel should increase the weighting of that channel. Comparing the proportions of illusory, visual-only, and auditory-only percepts will then index the relative strength of each weighting, indicating whether audiovisual speech relies more heavily on a familiar face or a familiar voice. I examine this question in the second chapter of this thesis.

1.5 Familiar Self-Speech

We recognize our own recorded voice as familiar (Kaplan, Aziz-Zadeh, Uddin, & Iacoboni, 2008), and when we see a videorecording of ourselves talking, we rely more heavily on our voice than our face (Aruffo & Shore, 2012). However, when we talk normally, we hear our own voice conducted through the bones of our head, as well as through the air (Békésy, 1949). When we hear a recording of ourselves, the bone-conduction effect does not occur, attenuating our perception of its higher frequencies (Shuster & Durrant, 2003) and making our recorded voice sound different from our "normal" voice. This difference is striking enough that, prior to the advent of convenient and inexpensive personal recording devices, listeners found it difficult to identify recordings of their own voices (e.g., Olivos, 1967). In the modern era, most people are exposed to their own recorded voice, and can identify a self-recording with near-ceiling

accuracy (e.g., Hughes & Nicholson, 2010), but we do not know if our experience with hearing self-voice is adequate to inculcate an advantage for recorded self-speech. Self-speech has not yet been tested to determine whether it conveys a familiar-talker advantage to word recognition. This question is addressed in the third chapter of this thesis.

1.6 The Role of Talker Identification

A familiar-talker advantage may be achieved through learning to identify a talker by name (Nygaard & Pisoni, 1998). From this result, it has been asserted that learning to identify a talker by name is a necessary condition for obtaining a familiar-talker advantage. However, a listener can learn to identify a talker without demonstrating a familiar-talker advantage for that talker's speech. This raises the possibility that learning to identify a talker by name is not a prerequisite for obtaining a familiar-talker advantage; rather, that the amount of learning necessary to achieve a familiar-talker advantage for a particular voice is more than sufficient to instill an association between the talker's voice and name. That is, gaining a familiar-talker advantage could also teach a listener to identify that talker, instead of the other way around.

Identifying a voice and recognizing its speech are separate processes, although interrelated. In normal listeners, judgments of acoustic and linguistic features proceed separately in parallel (Wood, 1974; Knösche, Lattner, Maess, Schauer, & Friederici, 2002). In brain-damaged listeners, it is possible to understand what a voice is saying while being unable to identify its talker (Van Lancker & Canter, 1982; Van Lancker &

Kreiman, 1987). These results suggest that normal listeners may be able to process familiar speech, and demonstrate a familiar-talker advantage, without needing to explicitly label a voice with its correct identity.

Curiously, although discussion of speech-processing models currently focuses on the relation between talker identity and speech recognition (cf. Goldinger, 1998), no investigations have yet been attempted in which listeners identify voices simultaneously with recognizing their speech contents. Finding an association, or lack of association, between correct identification of a talker's voice and accurate recognition of its speech, would contribute toward our understanding of the interdependence of vocal identity and speech. This is directly addressed in chapter three of the current thesis.

1.7 Goals

The goal of this thesis was to use familiar speech to explore how listeners relate a talker's identity to the speech he or she produces. Chapter 2 examined audiovisual speech perception, using the McGurk illusion to test whether a familiar face or voice contributed more strongly to optimal integration of auditory and visual channels. Chapter 3 tested self-speech, being the first investigation of whether recordings of self-voice convey a familiar-talker advantage to word recognition. Chapter 4 studied familiar auditory speech, investigating the interdependence of vocal identity and speech recognition by requiring listeners to identify voices while also transcribing their speech. The results of these experiments contribute to our understanding of how talker identity is encoded in memory and may be recalled for use in speech processing.

Chapter 2: Is a familiar face or a familiar voice more important to audiovisual speech?

2.1 Introduction

Speech is *audiovisual*. When someone talks to us, we "hear" both the words they say and the facial movements that produce that speech. Our minds automatically integrate both signals into a final speech percept, each modality modifying and informing the other. This is why speech is easier to understand when we can see the talker's face (Sumby & Pollack, 1954), or why it can be disorienting when vision and audition disagree, as with a foreign film dubbed into English (Dodd, 1977). When we hear speech in disorienting or degraded conditions, our minds resolve the problem by drawing more heavily from the more-reliable sensory channel (Massaro & Cohen, 1990). One way to make speech perception more reliable is to listen to a person we know, because familiar talkers are easier to understand (Nygaard & Pisoni, 1994). But which channel gives us a more-reliable speech signal: a familiar face, or a familiar voice?

2.1.1 Optimal integration

Faces and voices are presumed to have equal "weight" in audiovisual speech integration. Speech integration follows the principle of *optimal integration* (cf. Ernst & Bülthoff, 2004): all sensory inputs are "weighted" according to their perceived reliability, and a final speech percept can be predicted as a mathematical calculation of its combined weighted inputs (Massaro, 2004). When one channel is degraded, the clearer channel

makes a greater contribution to a final percept (MacDonald, Andersen, & Bachmann, 2000). When conditions are not degraded, then faces and voices contribute to speech perception as a function of their clarity relative to each other (Massaro, 2004). Equal contributions in either channel should therefore provide equally-weighted inputs, and an equally-clear face and voice should integrate equally in audiovisual speech.

But faces and voices are not necessarily equal in audiovisual integration. Faces receive greater weight from hearing-impaired listeners (Rouger, Fraysse, Deguine, & Barone, 2008; Schorr, Fox, van Wassenhove, & Knudsen, 2005), and voices receive greater weight from visually-impaired listeners (Putzar, Hötting, & Röder, 2010). Among people with normal hearing, individuals with schizophrenia or autism give greater weighting to voices (de Gelder et al., 2002; Smith & Bennetto, 2007). Speakers of different languages may give different weightings to voices or faces (e.g., Sekiyama & Tokhura, 1991), and their weightings can change depending on the sounds being presented (Massaro, Cohen, Gesi, Heredia, & Tsuzaki, 1993) or the perceived language of the talker (Hayashi & Sekiyama, 1998). In short, the reliability of each input channel, and, consequently, the weighting given to each sensory channel, depends not only on the clarity of its original signal but a listener's subjective judgments of whether the face or voice is more reliable.

2.1.2 Comparative contributions of faces and voices

In person recognition, faces are more reliable than voices. Faces are easier to recognize than voices (Yarmey, Yarmey, & Yarmey, 1994); faces can be identified swiftly and accurately, especially when they are in motion (Lander & Chuang, 2005), but

voices are identified more slowly and can be made unrecognizable by simple changes of tone (Saslove & Yarmey, 1980). If a talker's face and voice are both familiar, his or her face will be easier to recognize (Brédart, Barsics, & Hanley, 2009). It may be, therefore, that the subjective experience of a face's greater reliability in person identification will cause a familiar face to be weighted more heavily in audiovisual speech integration.

However, familiar faces and voices both convey speech-perception advantages. Speech is easier to understand when spoken by a familiar voice (Nygaard & Pisoni, 1994), and the more familiar the voice, the greater the advantage (Magnuson, Yamada, & Nusbaum, 1995). Lip-reading is more accurate, to a normally-hearing listener, when spoken by a familiar talker (Lander & Davies, 2008). Therefore, both a familiar face and voice could increase the reliability of their respective channels. On the other hand, being familiar with a talker's voice improves lip-reading accuracy (Sanchez, Dias, & Rosenblum, 2013), and familiar lip-movement improves auditory speech-recognition accuracy (Rosenblum, Miller, & Sanchez, 2007); so either of a familiar face or voice could increase reliability in both channels. To determine which is the greater contributor to audiovisual speech, and in which channel, the relative weights of a familiar face and voice must be measured.

2.1.3 The McGurk effect

The relative weights of auditory and visual channels may be measured with the *McGurk effect*. The McGurk effect (McGurk & MacDonald, 1976) is an audiovisual-speech illusion in which an auditory syllable (e.g., /ba/) is dubbed onto a conflicting visual syllable (e.g., /ga/), and a listener responds to this conflict by integrating the two

channels to perceive a "blended" syllable not actually present (e.g., /da/). The illusion is an ideal mechanism for testing audiovisual speech integration because it is automatic, mandatory, and robust. The illusion cannot be ignored, even when a listener has been instructed to do so (Massaro & Cohen, 1983). The illusion is not susceptible to spatial disunity (Jones & Munhall, 1997; Paré, Richler, ten Hove, & Munhall, 2003) or wide temporal asynchronies (Jones & Jarick, 2006; Munhall, Gribble, Sacco, & Ward, 1996), despite the fact that a listener can detect asynchrony (Soto-Faraco & Alsius, 2009; van Wassenhove, Grant, & Poeppel, 2007) and correctly indicate which channel was presented first (Soto-Faraco & Alsius, 2007; Vatakis & Spence, 2007). However, the illusion may be disrupted by a weighting imbalance. When either vision or audition is more reliable, the visual or auditory syllable will be reported instead of the illusion (Massaro, 1987). The McGurk effect can therefore be used to determine which of a familiar face or voice is perceived as more reliable in audiovisual speech processing.

2.1.4 Self-voice and self-face

In audiovisual self-speech processing, the voice is more reliable. Self-voice is a familiar voice (Nakamura et al. 2001), and self-face is a familiar face (Keyes, Brady, Reilly, & Foxe, 2010); indeed, self-face is identified more quickly and more accurately than other familiar faces (Keyes & Brady, 2010b). However, when self-speech was tested using the McGurk effect (Aruffo & Shore, 2012), such that listeners were presented with their own faces and voices mismatched to unfamiliar voices and faces, the illusion was disrupted only for stimuli featuring self-voice. Self-face stimuli supported as strong an illusion as did unfamiliar-talker stimuli. This result indicated that self-voice

contributes more strongly than self-face to audiovisual speech processing. This result, however, does not necessarily predict that a familiar face cannot affect audiovisual speech processing. Rather, it may simply mean that while we must hear our own voices when we speak, we rarely watch ourselves when we are talking.

2.1.5 Familiar talkers in audiovisual speech

We do not know which of a familiar face or voice contributes more strongly to audiovisual speech perception. Familiar audiovisual speech has been tested once before using the McGurk illusion (Walker, Bruce, & O'Malley, 1995), but the design of that experiment could not disambiguate faces and voices, because talkers were either all familiar or all unfamiliar. That is, the design was between-subjects, such that listeners were either familiar with all talkers or familiar with none. Listeners were presented with McGurk stimuli featuring either the same talker's voice and face (*matched identity*), or one talker's voice dubbed onto a second familiar talker's face (*mismatched identity*). The effect of familiarity was measured by comparing the proportion of illusions reported by each participant group (familiar versus unfamiliar). *Matched-identity* stimuli showed no effect of familiarity, supporting an equivalent proportion of illusions for familiar and unfamiliar talkers, whereas *mismatched-identity* stimuli supported fewer illusions from familiar talkers. However, because *mismatched-identity* stimuli presented a familiar talker in both channels, it is difficult to conclude how the illusion had been affected. When the illusion was not reported, listeners overwhelmingly reported the auditory percept (94%) rather than the visual percept (6%), suggesting that greater weight had been given to familiar voices—but if familiar voices were more strongly weighted than

familiar faces, familiar voices would have unbalanced *matched-identity* stimuli as well, and this did not occur. The reported effect of familiarity in this procedure may instead be due to listeners failing to integrate *mismatched-identity* stimuli. Listeners do not fail to integrate the McGurk illusion when unknown faces and voices are mismatched, as when an unfamiliar female voice is dubbed onto an unfamiliar male face, or vice versa (Green, Kuhl, Meltzoff, & Stevens, 1991); however, when both visual and auditory sources can be explicitly identified as incompatible, integration does not occur (Munhall, ten Hove, Brammer, & Paré, 2009; Saldaña & Rosenblum, 1993; Vatakis, Ghazanfar, & Spence, 2008). The presence of a distinctly-identifiable talker in each channel, therefore, may have averted integration, allowing listeners to attend selectively to the auditory channel. Therefore, we cannot conclude whether the results reported by Walker, Bruce, & O'Malley (1995) may be attributed to a familiar voice, a familiar face, or a failure to integrate two familiar talkers.

2.1.6 The present design

Our design presented the McGurk illusion with a condition where familiar talkers were dubbed onto unfamiliar talkers. If familiar speech is more reliable than unfamiliar speech, then the presence of a familiar talker in only one channel should create an imbalance in optimal weighting. Speech from a familiar talker should integrate with speech from talkers who cannot be explicitly identified (Aruffo & Shore, 2012). Therefore, presenting the McGurk illusion with a familiar talker in only one channel, and comparing its effect to that of unfamiliar talkers dubbed onto each other, should indicate whether familiar speech conveys greater weight to its sensory channel. If a familiar face

provides a speech-processing advantage, then mismatching it to an unfamiliar voice will produce greater weight in the visual channel, and listeners will perceive fewer illusions. If a familiar voice provides a speech-processing advantage, then mismatching it to an unfamiliar face will produce greater weight in the auditory channel, and listeners will perceive fewer illusions. Also, to test the relative contribution of a familiar face or voice, our design featured stimuli presenting the same familiar talker's face and voice in each channel. If either of a familiar face or voice contributed more strongly to audiovisual integration, then integration of a familiar talker's audiovisual speech would be unbalanced in favor of the more-reliable channel, and fewer illusions would be reported than from stimuli featuring unfamiliar talkers. The present design, therefore, by featuring familiar talkers in either the auditory channel, or the visual channel, or in both channels, should provide evidence as to whether a familiar face or voice makes a stronger contribution to audiovisual speech processing.

Training familiar voices

The strength of a familiar-talker effect is proportional to a listener's exposure to that talker (Magnuson, Yamada, & Nusbaum, 1995). To ensure equivalent levels of familiarity across participants, listeners in the present experiment received equivalent training to be familiar with a previously-unknown talker. Each participant was trained via prose narratives in one of three presentation types: *auditory*, *visual*, and *audiovisual*. For control, a fourth group remained *untrained*. Two classic prose narratives were selected: "Arthur the Rat" (Abercrombie, 1964) and "The Rainbow Passage" (Fairbanks, 1960). Each narrative was approximately two minutes long; passages of this length can

induce familiarity in either the visual or auditory domain via passive exposure (Lander & Davies, 2008; von Kriegstein et al., 2008). Although listeners differ in their ability to learn new voices (Nygaard & Pisoni, 1994), even "poor learners" can learn to identify a voice from fluent prose passages (Nygaard & Pisoni, 1998). Thus, listening to these prose narratives was expected to successfully familiarize a listener.

Stimulus design

To test whether talker familiarity had been successfully induced, non-illusory stimuli were featured. Non-illusory stimuli presented the same speech signal in both visual and auditory channels, allowing participants to provide *correct* or *incorrect* responses. If the training provided in the present procedure had successfully conveyed a familiar-talker speech-processing advantage, then this advantage could be measured as a proportion of correct answers given for familiar-talker stimuli versus unfamiliar-talker stimuli.

The McGurk illusion is famously robust, which could generate ceiling effects. Stimuli were therefore designed both to moderate the illusion and to facilitate listeners' recognition of familiar talkers. Each stimulus presented a vowel-consonant-vowel (VCV) nonword disyllable in which the vowel was always /a/. The vowel /a/ was selected because it supports a moderately-strong illusion (Green, Kuhl, & Meltzoff 1988; Hampson, Guenther, Cohen, & Nieto–Castanon 2003; Shigeno 2000). Nonwords were selected to avoid lexical effects (Brancazio, 2004), and presented in VCV configuration both to reduce word-onset influence (Barutchu et al., 2008) and to allow talker identification prior to the onset of the target consonant, as talkers can be identified from a

single sound (Craik & Kirsner 1974; Schweinberger, Herholz, & Sommer 1997; Beauchemin et al. 2006). Because simultaneous presentation of a face and voice can interfere with talker identification (Hughes & Nicholson 2010), each stimulus displayed a talker's face for approximately three seconds prior to verbal onset. The illusion was also moderated by listeners' method of response: because manual closed-response may produce a greater proportion of illusory responses (Colin, Radeau, & Deltenre, 2005), participants provided unconstrained verbal responses. Stimuli were therefore expected to support a moderately-strong illusion and—by providing participants ample time to identify familiar talkers—to facilitate familiar-talker effects.

Talker selection

Different talkers may not be equally intelligible. To minimize talker-specific effects, 19 different heterogeneous talkers were featured. Switching among different talkers from trial to trial was not expected to influence results. Switching talkers can interfere with perception of either auditory or visual speech (Mullennix, Pisoni, & Martin, 1989; Yakel, Rosenblum, & Fortier, 2000; Kaufmann & Schweinberger, 2005); however, the audiovisual McGurk illusion is not affected by switching talkers (Rosenblum & Yakel, 2001). The illusion is also not affected by female faces presented with male voices, or vice versa (Green, Kuhl, Meltzoff, & Stevens, 1991); therefore, every talker's voice was dubbed onto every other talker's face irrespective of talker sex. A drawback to featuring this many talkers was that, to achieve a desirable quantity of familiar-talker observations in a reasonable amount of time, a familiar talker must be presented more frequently than any of the unfamiliar talkers. This aspect of the design is

not a drawback to the extent that additional exposure may serve as further familiar-talker training; however, untrained listeners are likely to recognize that a familiar talker is being presented more frequently, and may gain some level of familiarity through mere exposure.

Listener selection

Listeners may not be equally susceptible to the McGurk illusion, regardless of familiarity level. Although listeners cannot deliberately choose to perceive or not perceive the illusion, listeners may not be equally likely to integrate audiovisual information (Rouger et al., 2007). The cause of such individual differences is currently under debate. Different listeners may have different levels of ability to integrate multimodal information (Grant, Walden, & Seitz, 1998; Grant, 2002). Alternatively, all listeners could be equally efficient at integration, but some listeners prefer one sensory modality over the other (Massaro & Cohen, 2000). Therefore, weaker integration could be explained by "auditory" and "visual" listeners giving undue weight to their favored channel, independently of the stimulus being presented (Schwartz, 2010). For the purpose of the present procedure, this debate reveals that a general population may comprise "good integrators" and "poor integrators." That is, given a randomly-selected group of observers, all with normal hearing, the group will not uniformly report high proportions of illusory percepts, but will instead produce a distribution that is negatively skewed (Grant & Seitz, 1998). If this distribution were to obtain from the current procedure, "good" and "poor" integrators could be examined by separating participants by the median of scores.

Speakers of different languages can be differently susceptible to the McGurk illusion, because a listener's first language can exert an influence on his or her relative weighting of auditory and visual information (Hardison, 1999). The present investigation was conducted on a university campus with a sizeable population of multilingual students; however, the influence of first language on audiovisual speech perception diminishes with experience (Navarra, Alsius, Velasco, Soto-Faraco, & Spence, 2010), and most university-level participants will have had substantial immersive experience with English. Each participant's first language was noted, and balanced among groups, but native language was not expected to produce a significant effect.

Order of presentation

A potential confound could arise from listeners, trained in one modality, learning to recognize their familiar talker in the other modality. People are highly skilled at matching moving faces with voices (Kamachi, Hill, Lander, & Vatikiotis–Bateson, 2003; Lachs & Pisoni, 2004). Identity matching is reliable even when the visual input is significantly degraded (Lachs & Pisoni, 2004b; Rosenblum, Smith, Nichols, Hale, & Lee, 2006) or the face and voice are not saying the same sentences (Lander, Hill, Kamachi, & Vatikiotis–Bateson, 2007). Most relevant to the present procedure is an experiment conducted by Daßler, Gottschlich, Itz, Knösing, and Temmerman (2008). Talkers were recorded speaking the phrase "You are what you think" (*"Du bist doch was Du denkst"*). Talkers' voices were then dubbed onto other talkers' faces, but each soundtrack was temporally adjusted to precisely follow the visual timing of its new face. These clips were then presented to listeners as a two-alternative forced-choice, in which listeners

selected which of the two clips featured the talker's original voice. Participants succeeded at this task with an accuracy rate of 92%. In the present procedure, talkers' voices were dubbed onto other faces and temporally adjusted to precisely follow their visual timings. It can therefore be expected that, over the course of an experimental session, trained participants will correctly judge which face (or voice) belongs to the voice (or face) with which they have been familiarized. However, regardless of listeners' ability to associate faces and voices, we do not know whether listeners' judgments of spoken syllables may be affected differently by faces and voices that either do or do not belong together. To ameliorate possible effects of associating faces and voices, stimuli were blocked. Stimuli featuring faces and voices that belonged together were blocked as *matched-identity*, and stimuli featuring two different talkers were blocked as *mismatched-identity*. *Matched-identity* blocks were always presented first.

Predictions

Whether a familiar-talker speech-processing advantage has been acquired, by listening to prose passages, will be tested by measuring participants' responses to non-illusory stimuli. A familiar-talker advantage would make non-illusory stimuli more intelligible, and would therefore be observed by a greater proportion of *correct* responses to stimuli featuring familiar talkers than to those featuring unfamiliar talkers.

If participants are differently susceptible to the McGurk illusion, participants presented with the same stimuli would report different proportions of illusory percepts. The McGurk illusion has historically been represented as mandatory and impenetrable, suggesting that participants would all report an equivalently-high proportion of illusions.

However, if some participants ("poor integrators") were less susceptible to perceiving the illusion, the proportions of illusions reported by participants would show a broader distribution. If a broader distribution is found, then the explanatory arguments on each side of the current debate will be evaluated by calculating participants' bias toward providing either auditory or visual responses. If individual differences can be best explained by personal biases toward one modality, such that observers' personal "weighting" of their preferred channel causes them to perceive that channel instead of the McGurk illusion, then a participant who shows "poor integration" should demonstrate a strong bias toward reporting a preferred mode, and a participant who shows "strong integration" should not. Otherwise, if individual differences may be better explained by differing levels of ability to integrate auditory and visual input, then participants who show "poor integration" are simply failing to integrate the illusion, and both "strong" and "poor" integrators would show similar patterns of auditory and visual responses.

A familiar face or voice may reduce audiovisual integration by increasing the "weight" in its respective sensory channel. Dubbing a familiar voice onto an unfamiliar face, or an unfamiliar voice onto a familiar face, would therefore create a weighting imbalance that would support fewer illusions. Therefore, it would be possible to measure the effect of a familiar face or voice by comparing the proportion of illusions reported for these stimuli to the proportion of illusions reported for unfamiliar faces and voices dubbed onto each other.

Familiar faces and voices may not contribute equally to multisensory integration. If their relative contributions are unequal, then integration would be imbalanced for

stimuli featuring the same familiar talker's face and voice together. Listeners would be less likely to perceive an illusion, and more likely to report the percept corresponding to the heavily-weighted channel. Alternatively, if familiar faces and voices do contribute equally, then a familiar talker may facilitate integration and thus enhance the McGurk illusion. Therefore, comparing the proportion of illusions reported for familiar stimuli versus unfamiliar stimuli, where the face and voice in each stimulus belong to the same talker, should indicate whether a familiar face or a familiar voice contributes more strongly to audiovisual speech processing.

2.2 Experiment

2.2.1 Method

Participants

165 participants (age 17–26 years) were recruited. Participants either spoke English as their native language (NL) or as a second language (ESL). Participants were assigned randomly to groups: 40 *auditory* (11 male, 29 female; 18 NL, 22 ESL), 40 *visual* (14 male, 25 female; 18 NL, 22 ESL), 39 *audiovisual* (21 male, 18 female; 18 NL, 21 ESL) and 46 *untrained* (10 male, 36 female; 46 NL). All participants gave informed consent to the procedures. All procedures complied with the tri-council ethics procedures in Canada as approved by the McMaster Research Ethics Board. All participants were McMaster University students who received course credit for their participation.

Stimuli

Audiovisual stimuli were recorded from 19 models (7 male, 12 female, age 18–45 years). To create training stimuli, models read each of two prose narratives: "Arthur the Rat" and "The Rainbow Passage." Each narrative was approximately two minutes long. To create testing stimuli, each model spoke five repetitions each of disyllables /aba/, /ada/, /aga/, /ala/, and /aða/, with a pause between syllables.

Models were recorded before a plain beige background in a sound-attenuated room. A digital video camera (JVC GZ-MG37U) and wireless lapel microphone (Shure PG185) were used. Two 60-watt lamps were placed at 45-degree angles to the participant's body, level with the participant's face, rendering the face clearly and fully visible. Video, framed to include the model's entire face, was recorded in 4:3 aspect ratio in MPEG-2 format (720 × 480 pixels, 8.5 Mbps). Audio was recorded separately in lossless WAV format (16-bit depth, 44100 samples/s). The audio and video of each session were synchronized to within 6 ms accuracy, using Adobe Premiere CS4, prior to editing, and saved to Windows Media format (WMV). All subsequent editing was accomplished using Adobe Premiere CS4. During the editing process, the audio of each stimulus was normalized to a perceived loudness of –15 dB using Adobe Audition 3.

Training stimuli were created from the prose-narrative readings. Three versions of each narrative were created: an *auditory* version in MP3 format (stereo, 128 kbps, 44100 samples/s), a silent *visual* version in which the soundtrack was removed, and an *audiovisual* version that presented both the visual and auditory signals. Thus 114 unique training stimuli were created (2 narratives × 3 versions × 19 participants). Not all

training stimuli were presented to each participant. These stimuli formed a pool from which the experimental procedure drew.

Testing stimuli were created from the disyllabic repetitions. Recordings were cut into four-second-long segments, each containing one disyllable. Each segment displayed a face gazing silently for approximately three seconds and then speaking one complete disyllable. In every stimulus, the consonant release was synchronized within an accuracy of 12 ms and the peak intensity of the initial vowel within an accuracy of 16 ms. All testing stimuli presented both a visual and an auditory signal.

Differences between stimuli were created by varying the talker identity and the disyllable presented in either modality. Talker identity was varied to create *identity-matched* and *identity-mismatched* stimuli: *identity-matched* stimuli presented a face and voice belonging to the same model, whereas *identity-mismatched* stimuli presented faces and voices belonging to different models. Each *identity-matched* stimulus was presented with its original video and audio. To create a corpus of *identity-mismatched* stimuli, each model's voice was dubbed onto every other model's face. Speaking-rate differences between talkers were accommodated by lengthening or shortening the pause between syllables. Disyllables were varied to create *congruent* and *incongruent* stimuli: *congruent* stimuli presented the same disyllable in both modalities, whereas *incongruent* stimuli presented the McGurk illusion (auditory /aba/ and visual /aga/). For every possible pairing of models' faces and voices, six stimuli were created: one *congruent* for each of the five recorded disyllables, plus one *incongruent* illusion. Because stimuli varied both talkers and disyllables presented, four classes of stimulus were generated:

congruent–matched, congruent–mismatched, incongruent–matched, and *incongruent–mismatched*. This resulted in a grand total of 2,622 unique audiovisual stimuli: 475 *congruent–matched* (5 disyllables × 5 repetitions × 19 models), 95 *incongruent–matched* (1 illusion × 5 repetitions × 19 models), 1,710 *congruent–mismatched* (5 disyllables × 18 voices × 19 faces) and 342 *incongruent–mismatched* (1 illusion × 18 voices × 19 faces). Each model's face and voice featured in 126 different stimuli. Not all stimuli were presented to every participant. These stimuli formed a pool from which the experimental procedure drew.

Procedure

Testing was conducted in a sound-attenuated room using a laptop computer (Gateway W6501) at 1280 × 800 resolution. Display brightness and volume were set to maximum. Audio was played from the computer's speakers situated directly below the screen. All participants were played one stimulus of a model speaking /ala/ to verify that stimuli were clearly audible but not uncomfortably loud.

Already-familiar models were removed from the testing procedure. Prior to testing, participants were shown a still-image gallery of all models and indicated any familiar faces. Participants were instructed to indicate "anyone you recognize by name." Models thus indicated were excluded from the procedure for that participant.

Each participant was pseudorandomly assigned to a particular model, such that all models were assigned with equivalent frequency across participants. Models' gender identity was not expected to be a factor (Green, Kuhl, Meltzoff, & Stevens 1991) and was not controlled. To induce familiarity, *trained* participants were each presented with prose

narratives ("Arthur the Rat" and "The Rainbow Passage"). One passage was presented before each of six testing blocks. The initial choice of narrative was randomly selected and then alternated with the other narrative; thus the two narratives were presented three times each during the training session. Narratives were presented in one of three modes: *auditory* (voice, black screen), *visual* (silent, dynamic face), or *audiovisual* (voice, dynamic face). Each participant received only one type of training. Participants were instructed to attend to the passage and were informed that they would not be required to remember the passages' content. Visual and audiovisual groups were instructed to "keep their eyes on the screen" during each passage. *Untrained* participants were each assigned to a "familiar" model, but were not informed of the model's significance nor presented with familiarizing passages.

In each trial, participants watched a single disyllabic token and verbally repeated what they had heard. Participants were instructed to maintain eye gaze on the screen and were informed that their looking away would produce invalid results. Participants underwent practice trials before commencing a testing session. In these practice trials, participants were reassured that there were "no wrong answers" and were instructed to respond quickly to avoid becoming confused. Responses were considered "quick" when spoken while the stimulus was still present on screen, and "delayed" when spoken after the stimulus had been removed from the screen. Practice continued until a participant had demonstrated three consecutive "quick" responses to incongruent stimuli. All participants successfully learned how to respond quickly.

Participants reported their own verbal responses. While a stimulus was presented, no screen elements other than the stimulus and experimental interface were visible. Three seconds subsequent to the stimulus' finish, after a participant had already spoken their response, five buttons were shown reading "ABA," "ADA," "AGA," "ALA," "ATHA," and "OTHER." Depending on the disyllable he or she had spoken, a participant either clicked one of the first five buttons, or typed a free response and clicked OTHER. Participants were informed that this report did not represent a "second chance" to change their response, but was meant to accurately report the disyllable that had been spoken. All responses were recorded using an omnidirectional microphone (Shure SM58) and were reviewed by random sampling to confirm the accuracy of participants' self-reports. Participants' self-reports generated the data analyzed here.

Six blocks of stimuli were presented at test. Three block types were featured: *matched*, *mismatched–familiar*, and *mismatched–unfamiliar*. Blocks contained different variations of stimuli as indicated in Table 1; a consequence of this design was that the familiar model's face and/or voice was present in 50% of all trials. Block-appropriate stimuli were selected randomly at runtime from the available pool, with the constraints that no two identical stimuli were presented consecutively, and that every unfamiliar model was presented with equivalent frequency across blocks. Each block type was presented twice during the session. Block order was constrained so that the first three blocks always included both *matched* blocks, and the latter three blocks always featured both *unfamiliar–mismatched* blocks; therefore, a pseudorandom order was determined by the positions of the *familiar–mismatched* blocks so that an equal number of participants

were presented with each possible ordering of all six blocks. Each of the six blocks presented 50 trials each, for a total of 216 *incongruent* and 84 *congruent* trials, or a grand total of 300 trials per testing session.

Table 2.1 Composition of stimulus blocks presented at test. Each set of 25 stimuli presented 18 *incongruent* and 7 *congruent* stimuli. Stimulus order was fully randomized within each block.

	Stimulus Block		
	Matched	Familiar–mismatched	Unfamiliar–mismatched
Stimuli Presented	25 familiar–matched	25 familiar-face	25 unfamiliar–mismatched
	25 unfamiliar–matched	25 familiar-voice	25 unfamiliar–mismatched

2.2.2 Results

Language

All between-groups tests described in these results were initially performed inclusive of the factor *language*, differentiating between participants whose native language was English and those whose native language was not English. However, *language* produced no effects or interactions in any test. For the sake of clarity, *language* as a factor has been omitted from the following results.

Familiar-talker advantage

To determine whether familiarity training had successfully induced a familiar-talker advantage, congruent (non-illusory) trials were tested for intelligibility. A congruent trial was *correct* when a participant accurately reported the disyllable presented, and *incorrect* otherwise. If familiarity training had conveyed an advantage, trained participants would report more correct answers than untrained participants. A mixed-design repeated-measures ANOVA (4 × 5) was performed on congruent trials using between-subjects factor *training* (auditory, visual, audiovisual, untrained) and within-subjects factor *stimulus identity* (familiar-match, unfamiliar-match, unfamiliar-mismatch, familiar-face, familiar-voice). Main effects were observed for *training*, $F(3, 161) = 4.67$, $p = .004$, and for *stimulus identity*, $F(4, 644) = 9.93$, $p < .001$; these factors did not interact. To interpret the effect of *training*, post-hoc Dunnett two-tailed t-tests were performed, using *untrained* as the control group (Figure 2.1). These tests indicated that each trained group had given a greater proportion of *correct* answers than did untrained participants ($p \leq .015$). The main effect of *stimulus identity* suggested that

increased accuracy for trained groups might be attributable to familiar stimuli; therefore, planned comparisons were performed on accuracy scores using Wilcoxon signed-rank tests ($\alpha = .05$). Non-parametric tests were used because accuracy scores were not normally distributed (Figure 2.2), $W(165) = .83$, $p < .001$, having a skewness of –2.10 (SE = 0.19) and kurtosis of 6.84 (SE = 0.38). These tests showed greater accuracy for all familiar stimuli: *familiar–matched* versus *unfamiliar–matched*, $Z = –3.75$, $p < .001$; for *familiar-face* versus *unfamiliar–mismatched*, $Z = –2.81$, $p = .005$, and for *familiar-voice* versus *unfamiliar–mismatched*, $Z = -4.95$, $p < .001$. Although the interaction between *training* and *stimulus identity* was not significant, all trained-listener groups demonstrated greater word-recognition accuracy than untrained listeners, and all three familiar-stimulus types were more intelligible than unfamiliar stimuli. These results support our *a priori* understanding of a familiar-talker advantage.

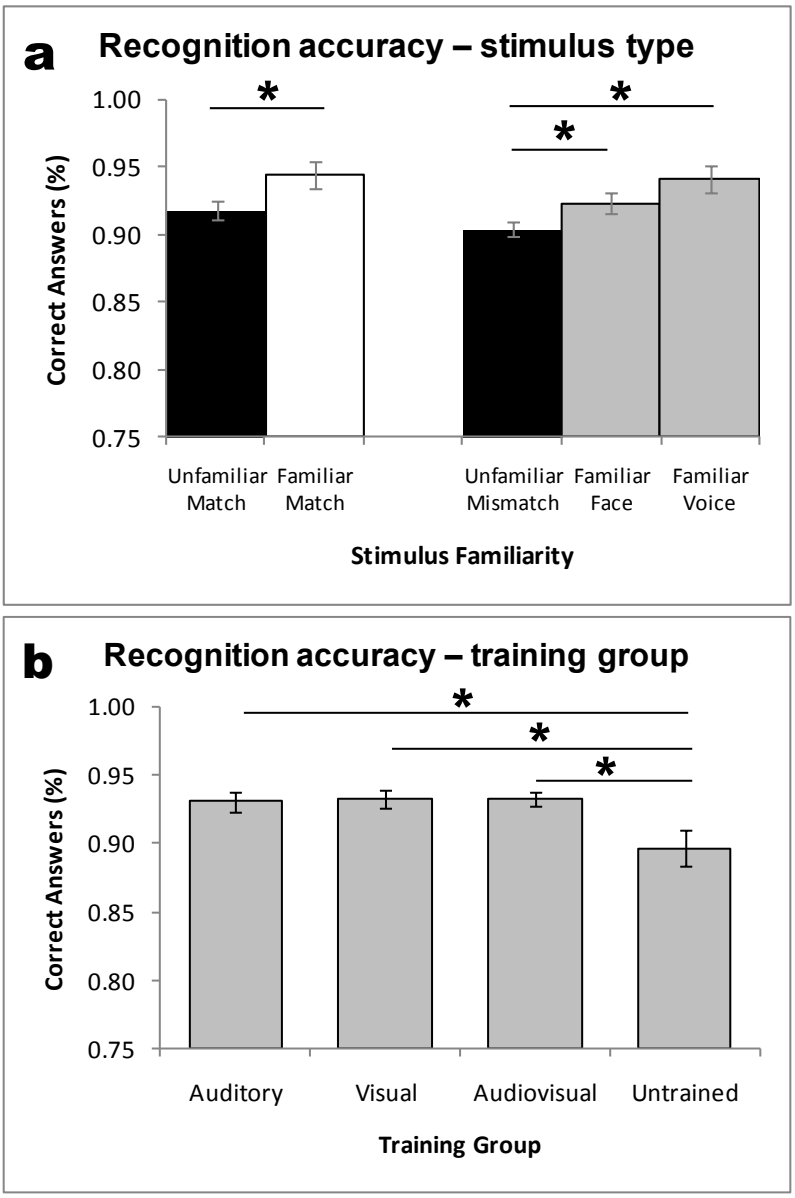

Figure 2.1 Word-recognition accuracy for congruent (non-illusory) trials. Significant differences are indicated by asterisks. Error bars are standard error of the mean corrected for within-subjects design (a) or standard error of the mean (b).

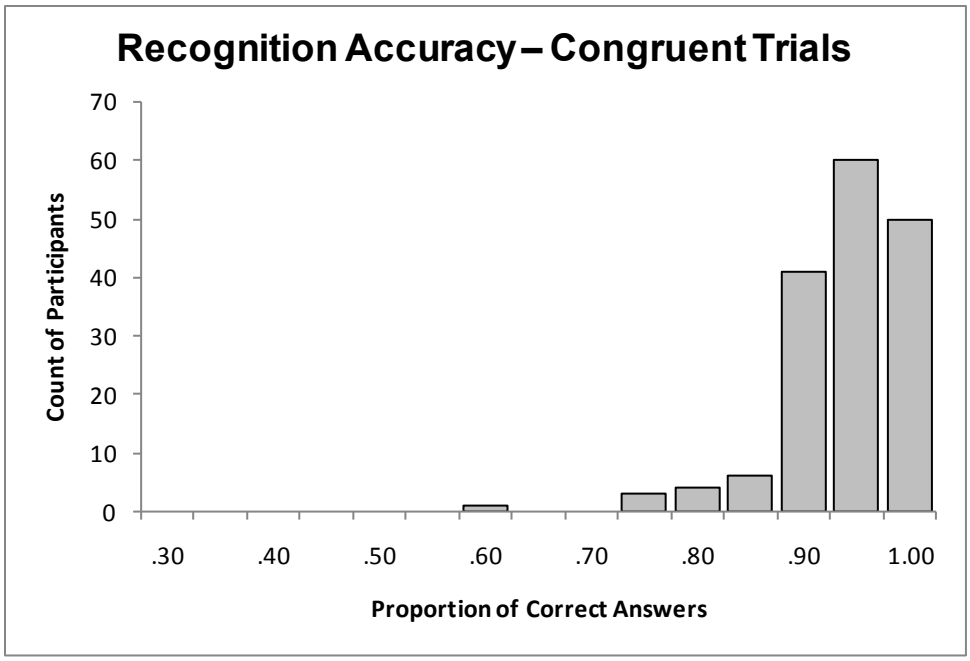

Figure 2.2 Distribution of recognition-accuracy scores for all participants.

Based on our *a priori* understanding of a familiar-talker advantage, and the familiarity advantages already observed (Figure 2.2), trained and untrained groups were independently tested for an effect of *stimulus identity*, using Friedman tests ($\alpha = .0125$). All trained groups showed a significant effect, and untrained did not: *auditory*, $\chi^2(4) = 61.40$, $p < .001$; *visual*, $\chi^2(4) = 12.87$, $p = .012$; *audiovisual*, $\chi^2(4) = 25.39$, $p < .001$; *untrained*, $\chi^2(4) = 6.24$, ns. Trained groups did not differ from each other; a Kruskal–Wallis test performed on *training*, excluding *untrained* participants, showed no differences between groups for any level of *stimulus identity*. Trained groups were therefore collapsed for further intelligibility testing. To determine whether trained groups had perceived familiar stimuli more accurately, planned comparisons were performed between familiar and unfamiliar stimuli, using Wilcoxon signed-rank tests ($\alpha = .05$). These tests indicated that trained participants gave more correct answers for all three familiar-stimulus types than they did for unfamiliar stimuli ($p < .001$), whereas untrained participants' accuracy did not differ among stimuli (Figure 2.3). Given the results of these tests, and our *a priori* understanding of a familiar-talker advantage, the speech-perception advantage demonstrated by trained participants appears to have been driven by familiar stimuli; this result would confirm that all types of training did successfully induce a familiar-talker speech-processing advantage.

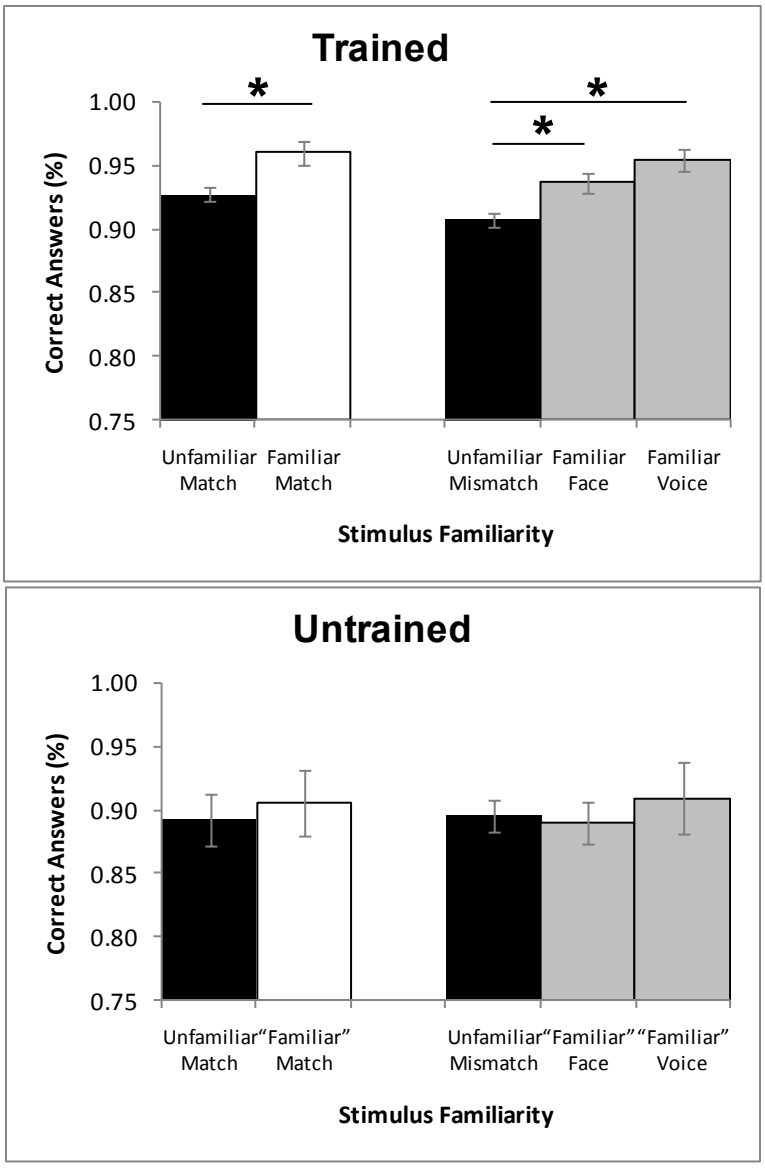

Figure 2.3 Syllable-recognition accuracy for congruent (non-illusory) trials, grouped by *training*. Significant differences are indicated by asterisks. Error bars are standard error of the mean corrected for within-subjects design. N.B.: Untrained participants were not familiarized with their assigned "familiar" talkers.

Due to our *a priori* expectation that different participants could exhibit different predispositions toward perceiving the McGurk illusion, the distribution for incongruent-trial data was tested for normality. Incongruent trials were coded as follows: *visual* when a participant accurately reported the visual disyllable; *auditory* when the auditory disyllable was reported; and *illusory* otherwise. Our primary measure was the proportion of *illusory* reports from each participant. These data were not normally distributed, $W = .80$, $N = 165$, $p < .001$, with skewness of -1.44 (SE = .19) and kurtosis of 1.18 (SE = .38) (Figure 2.4). Given the near-ceiling performance of the top half of the distribution, we chose to separate the data into two groups, divided at the median proportion of illusions (.91). This method appeared to capture the interesting aspects of the data. Participants scoring below the median were classified as *low integrators*, and participants scoring above the median were classified as *high integrators*. Low integrators comprised 81 participants: 22 auditory, 17 visual, 17 audiovisual, and 25 untrained. High integrators comprised 84 participants: 18 auditory, 23 visual, 22 audiovisual, and 21 untrained. Tests of familiarity effects were performed separately on low integrators and high integrators.

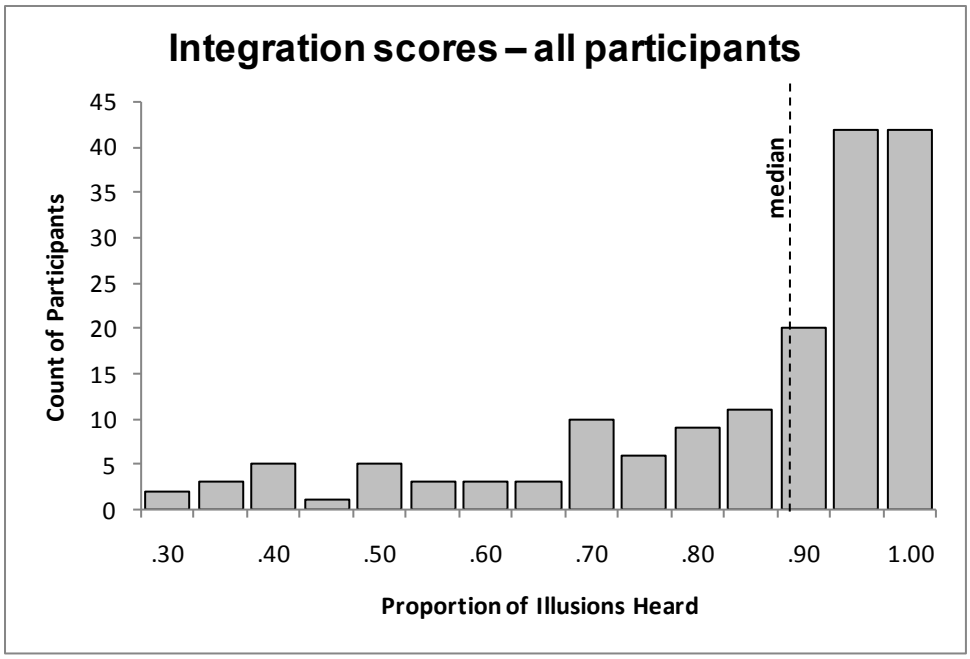

Figure 2.4. Distribution of integration scores across participants. Median value is indicated with a dashed line. Scores represent proportion of illusions reported in incongruent trials.

Low integrators

To determine whether talker familiarity had affected low integrators' perception of the McGurk illusion, a mixed-design ANOVA (4 × 5) was performed on incongruent-trial data using factors *training* (auditory, visual, audiovisual, untrained) and *stimulus identity* (unfamiliar–matched, familiar–matched, unfamiliar–mismatched, familiar-face, familiar-voice). Mauchly's test indicated that sphericity was violated, $\chi^2(9) = 78.71$, $p < .001$, so Greenhouse-Geisser correction was applied. A main effect was observed for *stimulus identity*, $F(3, 204) = 7.31$, $p < .001$, but not for *training*, $F(3, 77) = 0.68$, n.s. The two factors did not interact. To support that these results were not an error due to a non-normal distribution, non-parametric tests were performed. To support that *training* had not produced an effect, a Kruskal–Wallis test was performed on low-integrator data, and no differences were found between groups for any level of *stimulus identity*. To support the main effect of *stimulus identity*, a Friedman test was performed, and this test confirmed the main effect, $\chi^2(4) = 14.62$, N = 81, $p = .006$. To determine which stimuli had driven the effect of *stimulus identity*, planned comparisons were performed using Wilcoxon signed-ranks tests ($\alpha = .05$). Because no effect of *training* had been found, these tests were performed on all data, across *training* groups. These tests (Figure 2.5) showed that fewer illusions had been perceived for *familiar–matched* than *unfamiliar–matched* stimuli, $Z = –2.72$, $p = .007$, but that *unfamiliar–mismatched* was not different from either *familiar-face*, $Z = –0.46$, n.s., or *familiar-voice*, $Z = –1.75$, n.s. Therefore, for low integrators, talker familiarity reduced the proportion of *familiar–matched* illusions perceived, but this effect could not be attributed to training.

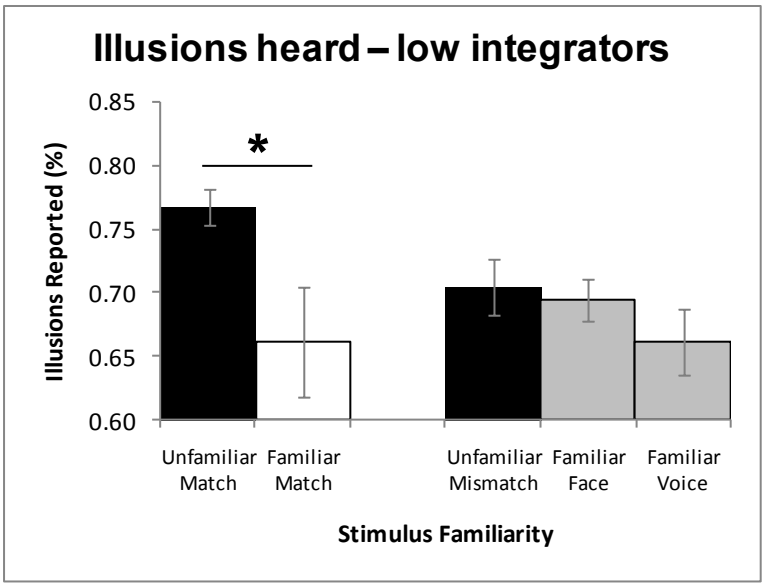

Figure 2.5 Proportion of illusions heard by low integrators for each stimulus type. Note that data are collapsed across all *training* groups (including untrained participants) because there was no interaction with this factor (but see Figure 2.6). Significant differences are indicated with asterisks. Error bars are standard error of the mean corrected for within-subjects design.

The failure of either *familiar-face* or *familiar-voice* stimuli to show an effect was surprising. Because *familiar–matched* stimuli supported fewer illusions, one of *familiar-face* or *familiar-voice* had, presumably, been more heavily weighted than the other. Therefore, the heavier channel should also convey greater weighting when an unfamiliar talker was presented in the opposite channel, and affect the proportion of illusions supported by stimuli featuring that stronger channel, meaning that either *familiar-face* or *familiar-voice* should support a different proportion of illusions than *unfamiliar–mismatched*. However, no such difference was observed. This lack of difference may be due to the lack of effect shown by *training*. The main effect of *stimulus identity* did not interact with *training*, which meant that the effect included data from both *trained* and *untrained* participants, and *untrained* participants were not trained to recognize their "familiar" talker. The inclusion of untrained participants' data may, therefore, have obscured an effect of training on *familiar-face* or *familiar-voice* stimuli. To explore this speculation, planned comparisons were performed on trained-participant data, using Wilcoxon signed-rank tests. These tests examined differences between familiar and unfamiliar stimuli as a result of training. Fewer illusions were supported by *familiar–matched* stimuli than *unfamiliar–matched*, $Z = -2.45$, $p = .014$, and by *familiar-voice* versus *unfamiliar–mismatched*, $Z = -2.07$, $p = .038$, but *familiar-face* was not different from *unfamiliar–mismatched*, $Z = -1.05$, ns (Figure 2.6a). Similar tests performed on *untrained* participants' data did not show significance (Figure 2.6b), but this may be due to lack of power. Tests on trained participants' data therefore support the speculation that

the weaker illusion supported by *familiar–matched* stimuli may have been attributable to the stronger contribution of a familiar voice.

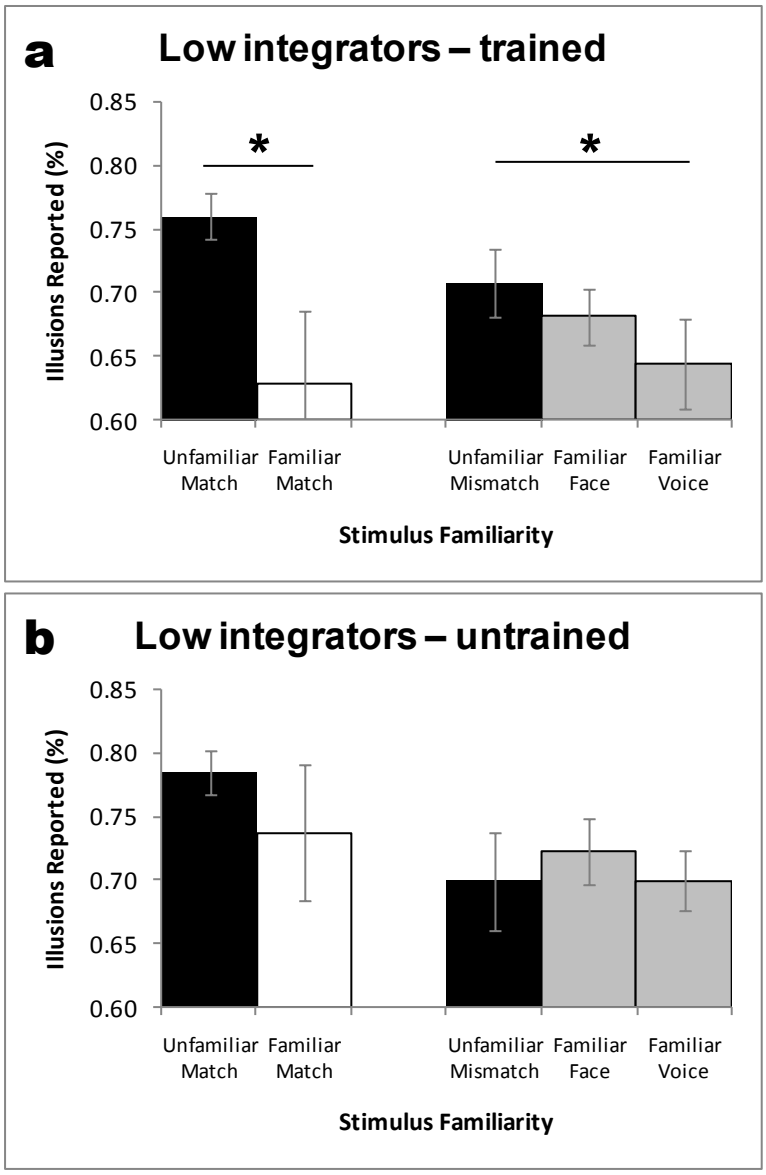

Figure 2.6 Illusions perceived by low integrators, grouped as *trained* or *untrained*. Significant differences are indicated by asterisks. Error bars are standard error of the mean corrected for within-subjects design.

High integrators

To determine whether talker familiarity had affected high integrators' perception of the McGurk illusion, a mixed-design ANOVA (4 × 5) was performed on incongruent-trial data using factors *training* (auditory, visual, audiovisual, untrained) and *stimulus identity* (unfamiliar–matched, familiar–matched, unfamiliar–mismatched, familiar-face, familiar-voice). Mauchly's test indicated that sphericity was violated, $\chi^2(9) = 32.08$, $p < .001$, so Greenhouse-Geisser correction was applied. Main effects were observed both for *stimulus identity*, $F(3, 265) = 6.96$, $p < .001$, and *training*, $F(3, 80) = 5.29$, $p = .002$; these factors did not interact. To interpret the main effect of *training*, post-hoc Dunnett two-tailed t-tests were performed, using *untrained* as the control group. These tests indicated that more illusions were reported by those trained *visually*, $t(42) = 3.75$, $p = .001$, or *audiovisually*, $t(41) = 3.34$, $p = .039$, than were reported by *untrained* participants, but that *auditory* training was not different from *untrained* (Figure 2.7). To support that these results were not due to an error arising from non-normal distribution, two non-parametric tests were performed. Firstly, to support the main effect of *stimulus identity*, a related-samples Friedman's test was performed, and this test showed significance, $\chi^2(4) = 31.13$, N = 84, $p < .001$. Secondly, to support the main effect of *training*, a Kruskal–Wallis test was performed on *training* data for different levels of *stimulus identity*. This test showed differences among training groups for *familiar–matched* stimuli, $\chi^2(3) = 13.81$, N = 84, $p = .003$, and for *familiar-voice* stimuli, $\chi^2(3) = 11.73$, N = 84, $p = .008$. To determine whether these two stimulus types could have driven the effect of *stimulus identity*, planned comparisons were performed between

matched trials and among *mismatched* trials, Wilcoxon signed-ranks tests ($\alpha = .05$). These tests showed that *familiar–matched* stimuli supported more illusions than *familiar-mismatched* stimuli, $Z = -3.55$, $p < .001$, and *familiar-voice* stimuli supported more illusions than *unfamiliar–mismatched* stimuli, $Z = -2.63$, $p = .009$. *Familiar-face* was not different from *familiar-mismatched*, $Z = -0.71$, n.s. Therefore, for high integrators, *visual* and *audiovisual* training increased the proportion of illusions perceived, and more illusions were perceived for *familiar–matched* and *familiar-voice*. Although the interaction was not significant, these results are consistent with our *a priori* expectation that training would affect perception of familiar stimuli.

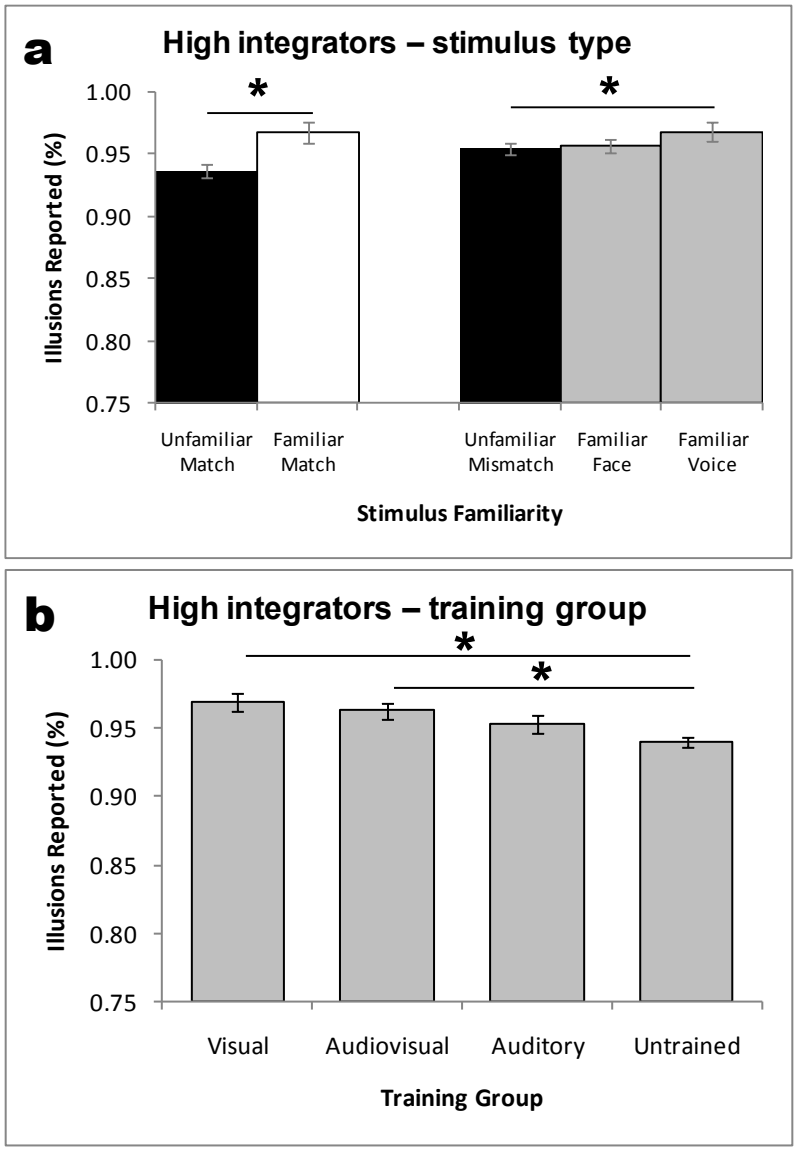

Figure 2.7 Proportion of illusions heard by high integrators. Significant differences are indicated with asterisks. Error bars are standard error of the mean corrected for within-subjects design (a) or standard error of the mean (b).

Auditory response bias

To determine whether participants were biased toward either the auditory or visual mode, data were analyzed from incongruent trials where a *unimodal* response—auditory or visual, rather than illusory—had been given. Because our interest was in comparing the relative proportions of auditory and visual responses from each participant, an *auditory response bias* score was calculated for each participant. This was accomplished by calculating a participant's proportion of auditory scores versus visual scores, $(a - v) \div (a + v)$, and then normalizing each participant's score across all participants, as (*participant mean – grand mean*) ÷ *standard deviation*, or (X – M) ÷ SD. A positive *auditory response bias* score represented a bias toward giving auditory responses, and a negative score represented a bias toward giving visual responses. A zero score would represent no significant bias for either modality.

Three hypotheses were tested by analyzing *auditory response bias* scores: one, whether listeners' overall performance as "high" or "low" integrators could be explained by modal response bias; two, whether the effect of a familiar voice could be explained by greater weighting in the auditory channel; and three, if listeners' weighting of faces or voices had been affected by training mode.

If listeners' overall performance as "high" or "low" integrators could be explained by modal response bias, then low integrators would show strong biases toward their preferred modalities, whereas high integrators would show weaker bias. Biases would be observable in the distribution of *auditory response bias* scores for either group. If low integrators were more strongly biased, their scores would appear polarized, with

participants tending toward either extreme of auditory or visual bias. If high integrators were less susceptible to bias, their scores would tend toward a neutral (zero) value and more closely resemble a normal distribution. Shapiro–Wilk tests performed on each group indicated that data were not normally distributed either for low integrators, $W(81) = .86$, $p < .001$, skewness = –0.60 (SE = .27), kurtosis = –1.17 (SE = .53), or for high integrators, $W(84) = .91$, $p < .001$, skewness = –0.07 (SE = .26), kurtosis = –1.36 (SE = .52). Moreover, both high and low integrators were susceptible to strong biases (Figure 2.8). Therefore, it is unlikely that listeners' overall integration performance may be attributable to modal response bias.

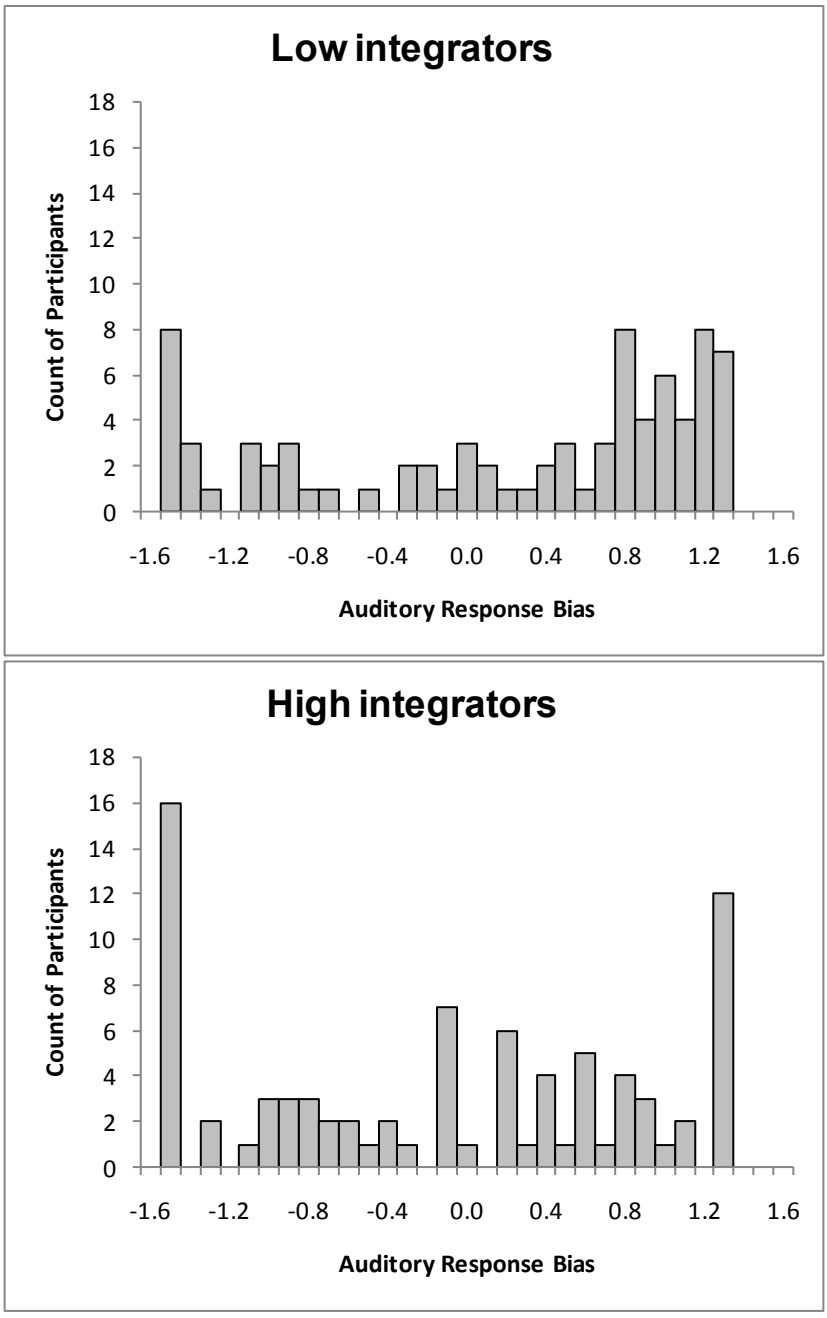

Figure 2.8 Distributions of *auditory response bias* scores for low integrators (A) and high integrators (B). Negative scores represent bias toward the visual channel.

To determine whether any of the factors in the current design were responsible for modal biases, a mixed-design ANOVA (2 × 4 × 5) was performed on *auditory response bias* data, using factors *integration level* (low, high), *training* (auditory, visual, audiovisual, untrained), and *stimulus identity* (unfamiliar–matched, familiar–matched, unfamiliar–mismatched, familiar-face, familiar-voice). *Integration level* produced no effects or interactions. To support that this result was not due to a non-normal distribution of data, three tests were performed. To support that *integration level* had produced no main effect, a Mann–Whitney paired comparison was performed on *integration level* (low, high) for overall *auditory response bias* scores; to support that *integration level* had not interacted with *stimulus identity*, a Kruskal–Wallis test was performed on *integration level* across levels of *stimulus identity*; to support that *integration level* had not interacted with *training*, Mann–Whitney tests were performed to compare *integration level* for each level of *training*. None of these tests showed any significant result. *Integration level* was therefore omitted as a factor from further testing.

To analyze the latter two hypotheses—whether a familiar voice conveyed greater weighting to the auditory channel, and whether modal weightings had been influenced by training—a mixed-design ANOVA (4 × 5) was performed using factors *training* (auditory, visual, audiovisual, untrained), and *stimulus identity* (unfamiliar–matched, familiar–matched, unfamiliar–mismatched, familiar-face, familiar-voice). Mauchly's test indicated that sphericity was violated, $\chi^2(9) = 25.40$, $p = .003$, so Greenhouse-Geisser correction was applied. No main effect of *stimulus identity* was observed, $F(4, 593) = 0.01$, ns, but there was a main effect of *training*, $F(3, 161) = 9.02$, $p < .001$, and the two

factors interacted, $F(11, 593) = 2.00$, $p = .026$. To examine the main effect of *training*, post-hoc comparisons were performed between each trained group and untrained participants, using Mann–Whitney tests ($\alpha = .0125$). Untrained listeners had demonstrated an overall auditory bias; this was different from the *visual* group, $Z = -3.30$, $p < .001$, and *audiovisual* group, $Z = -3.32$, $p < .001$, but not *auditory*, $Z = -1.18$, ns. The interaction was supported by a Kruskal–Wallis test performed on *training* for different levels of *stimulus identity* (Figure 2.9); this test showed significant between-group differences at all levels except *familiar-face* ($p \leq .011$). To determine which stimulus types had been affected by training, paired comparisons were performed on each stimulus type, comparing *visual* to *untrained* and *audiovisual* to *untrained* using Mann–Whitney tests ($\alpha = .01$). *Untrained* participants demonstrated auditory bias for all stimulus types; therefore, finding a significant between-groups difference for any stimulus type would indicate that training had induced a visual bias for that stimulus type. *Visual* training induced a visual bias for *unfamiliar–matched*, $Z = -5.10$, $p < .001$, and *familiar–matched*, $Z = -3.97$, $p < .001$, but not for any mismatched stimuli. *Audiovisual* training induced a visual bias for all stimulus types but *familiar-face*: *unfamiliar–matched*, $Z = -4.24$, $p < .001$; *familiar–matched*, $Z = -3.74$, $p < .001$; *familiar–mismatched*, $Z = -3.20$, $p = .001$; and *familiar-voice*, $Z = -3.23$, $p = .001$. Together, these tests indicated that participants' tendencies to provide auditory or visual responses to incongruent stimuli were not driven by weighting from familiar voices or faces, but by the training modes to which participants had been exposed.

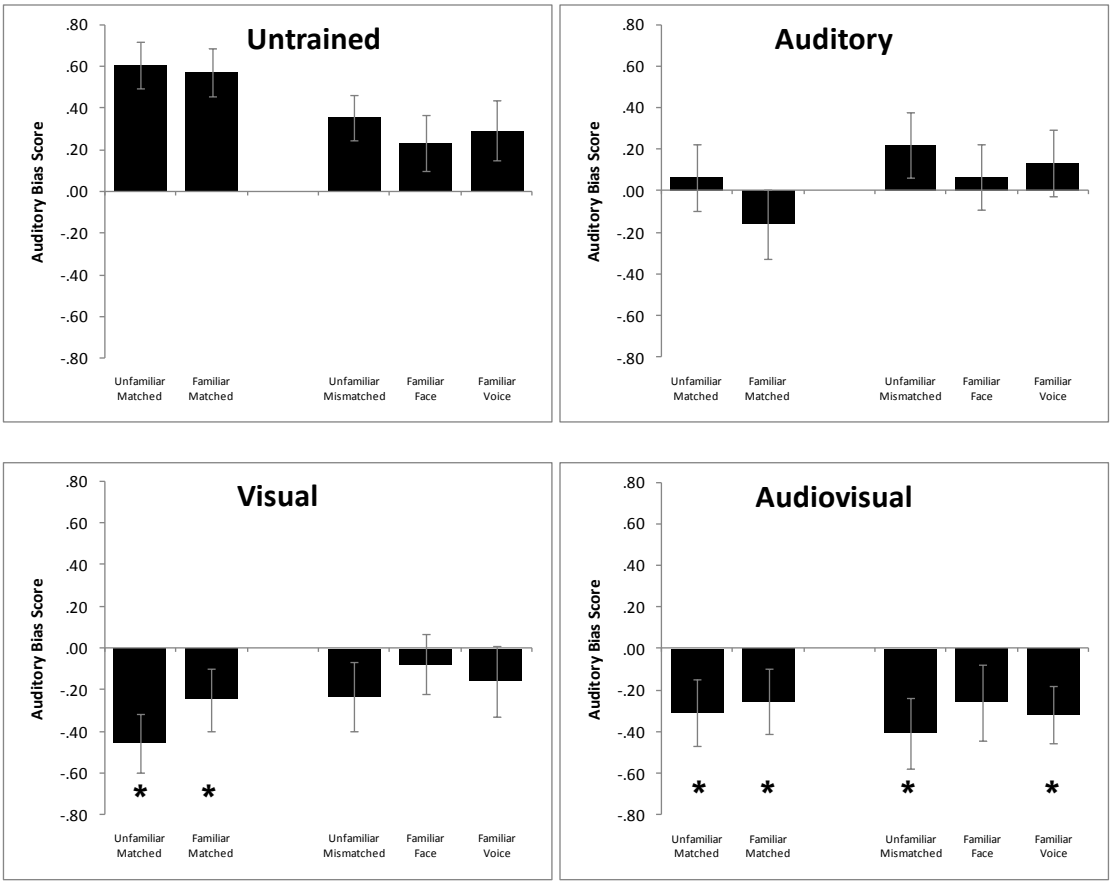

Figure 2.9 Auditory response bias scores for all training groups, showing a two-way interaction between *training* and *stimulus identity*. Visual biases induced by training are marked with asterisks. Error bars are standard error of the mean.

2.3 Discussion

Participants were familiarized with a talker and then presented with the McGurk illusion, spoken by that familiar talker and others. Participants were not equally susceptible to the illusion. Regardless of training, half of all participants integrated a high proportion of illusions, and the other half a low proportion of illusions. High integrators and low integrators showed opposite effects of familiarity: familiar talkers caused low integrators to perceive fewer illusions and high integrators to perceive more. High integrators were affected by familiar-face training and not by familiar-voice training, and low integrators were less affected by training than by the frequency with which familiar talkers were presented. However, familiar-talker effects for all participants appeared to be driven by presentation of a familiar voice, and not a familiar face.

When illusions were not perceived, participants' responses were not consistent with those predicted by optimal integration. A participant's bias to report auditory or visual percepts was not determined by the presence of a familiar voice or face in a given stimulus, but instead was determined by the sensory mode in which the participant had been trained. These results suggest that familiar-talker effects in audiovisual integration may not be fully explained by optimal integration, and that different listeners may possess different strategies for integrating audiovisual speech.

2.3.1 Levels of integration

Listeners in the present experiment did not integrate audiovisual speech at equal levels. Half the participants were *low integrators*, perceiving between 28% and 90% of

the illusions presented to them. The other half were *high integrators*, perceiving 91% or more. That listeners may integrate at different levels is not a new observation (Grant & Seitz, 1998), but the reason for individual differences is not yet known. Two hypotheses have been advanced to explain between-listener differences. The first hypothesis asserts that listeners are not equally skilled at integrating audiovisual speech, and will therefore be more or less successful at doing so independently of the speech signal's quality (Grant, 2002; Grant, Walden, & Seitz, 1998). The second hypothesis contends that all listeners are equally skilled at audiovisual integration, and that "success" at integration is wholly dependent on the quality of information provided in each sensory mode (Massaro & Cohen, 2000). The present experiment contributes new evidence to this conversation by showing not merely that listeners integrated at different levels, but that low and high integrators responded differently to familiar talkers. The effect of familiarity occurred in opposite directions: low integrators integrated less, and high integrators integrated more. This result might have been caused by differing integration ability, or might instead be due to different strategies for evaluating information.

2.3.2 Mechanisms of integration

The present results do not argue for differences in overall integration ability. If the mechanism driving the current results were listeners being either "good" or "poor" at audiovisual integration, then familiar talkers made "good" integrators better and "poor" integrators worse. However, there is no precedent for familiar talkers conveying a disadvantage to speech processing. Normally, familiar speech is recognized more quickly and more accurately (Craik & Kirsner, 1974; Nygaard & Pisoni, 1994); this was

also the case here, as all trained listeners recognized familiar talkers' non-illusory speech more accurately. "Poor" integrators, therefore, did learn to perceive familiar speech more accurately—and historical evidence does not suggest that perceiving speech more accurately can cause a deficit in audiovisual integration. If gaining a speech-processing advantage caused some listeners to integrate more, and others less, different listeners may have used that same advantage differently.

Listeners did appear to exhibit different integration strategies. Listeners may be "more auditory" or "more visual" in their orientation to audiovisual speech (Schwartz, 2010); the present results could be explained by low integrators being "more auditory" and high integrators "more visual." Low integrators may have learned to perceive a familiar talker's voice, even from visual training (Rosenblum, Miller, & Sanchez, 2007); then, upon presentation of a familiar voice, low integrators would weight the auditory channel more heavily and perceive fewer illusions. High integrators, who were affected only by visual or audiovisual training, may have learned to perceive a familiar talker's facial speech, and learned to associate that face with its matching voice (Daßler, Gottschlich, Itz, Knösing, & Temmerman, 2008); therefore, upon presentation of a familiar voice, high integrators would have used their memory of the familiar face to simulate the familiar face, and match it to the familiar voice, thus performing more efficient integration and perceiving more illusions (von Kriegstein et al., 2008). If listeners performed these strategies generally, for familiar and unfamiliar stimuli, this would mean that "low integrators" focused on the auditory channel to the exclusion of vision, and "high integrators" focused on the visual channel to the benefit of audition.

Listeners' use of these different strategies does not necessarily represent different levels of ability to perform audiovisual integration; however, it does suggest that "visual" listeners work toward integration and "auditory" listeners work against it.

2.3.3 Optimal integration and modal bias

Listeners' responses were not consistent with optimal integration. When illusions were not perceived, optimal integration would predict that the choice of response would be determined by which channel was more heavily weighted due to its perceived reliability. If low integrators were "more auditory," and high integrators "more visual," each group could be expected to weight their preferred channel more heavily; thus, among trials reported as non-illusory, each low integrator should be biased toward providing auditory responses, and each high integrator should be biased toward visual responses. Moreover, if familiar-talker effects were driven by a familiar voice, then low integrators' heavier weighting in the auditory channel should produce more auditory responses for familiar-voice stimuli. Neither of these predictions were borne out. Instead, a listener's bias toward visual or auditory responses was dependent on the type of training received. Untrained listeners were biased to give auditory responses. Trained listeners showed either a reduced auditory bias or a significant visual bias. These results raise the possibility that trained listeners' auditory and visual responses to illusory stimuli may not be due to optimal integration, but failed integration, after which listeners were able to selectively attend to the modality toward which they had been oriented. Integration could not have failed due to a different distinctly-recognizable talker in both channels (Walker, Bruce, & O'Malley, 1995), because listeners in the present experiment

were familiar with only one talker. However, the McGurk effect is weakened when listeners are distracted, either by a sudden effect or a secondary task (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Alsius, Navarra, & Soto-Faraco, 2007; Tiipana, Andersen, & Sams, 2010). Because each stimulus began by presenting three seconds of an unfamiliar face, trained listeners may have found that suddenly hearing a familiar voice, and identifying that voice, was sufficiently distracting to prevent integration. Additionally, listeners who expect audiovisual incongruity are less likely to integrate the McGurk illusion (Nahorna, Berthommier, & Schwartz, 2012). A familiar-talker processing advantage may have made the incongruity of familiar stimuli more obvious, and thereby increased trained listeners' overall sensitivity to whether any stimulus' incongruent auditory and visual signals should be bound together. In any event, the failure of trained listeners' modal biases to conform to those predicted by optimal integration suggests that the effects of training in audiovisual speech integration may not be entirely explained by optimal-integration mechanisms.

Neither a familiar face nor a familiar voice automatically "contributed more" to optimal integration. Rather, the contribution of familiar speech depended on the channel from which a listener was more capable of learning. Low integrators were more affected by exposure to a familiar voice, and high integrators were more affected by training with a familiar face. However, it can be argued that the current results, consistent with the findings from self-talkers (Aruffo & Shore, 2012), show that a familiar voice was more important to audiovisual integration. Although statistical power was not great enough to draw strong conclusions, familiar-talker effects appeared to be activated by presentation

of a familiar voice, and not a familiar face. Low integrators penetrated the illusion presented by familiar-voice stimuli, but not familiar-face; and high integrators were better able to integrate familiar voices, but not familiar faces, despite having been trained visually. This latter result strongly suggests that listeners, despite "being visual" when learning familiar speech, nonetheless gave precedence to auditory input when listening.

2.3.4 Weaknesses of the present design

The present design made the effects of training difficult to interpret. Individual differences in overall integration levels were anticipated, and were accommodated by separating listeners into high integrators and low integrators and testing each population separately. That this separation should cause the reduction in power that it did was not anticipated. Consequently, for both high and low integrators, tests performed among different training groups showed effects of training and of familiar talkers, and simple-effects tests performed on each training group supported each of our *a priori* expectations—that trained and untrained participants would behave differently, and that different training would produce different responses—but the interactions supporting these expectations did not reach significance. If the present design were altered, both by reducing the quantity of talkers and by eliminating two of the four training groups, tests should then have enough power to interpret *a priori* expectations.

Differences among low-integrator groups were obscured by an effect of exposure during testing. Because the familiar talker was presented in half of all trials, low integrators who had not received training were nonetheless able to recognize and become familiar with that talker. Untrained listeners thereby gained an effect of familiarity which

could not be easily distinguished from the effect induced by training. This exposure effect could be eliminated by using four talkers rather than nineteen. Using only four talkers would allow all talkers to be presented with equal frequency during training and yet still provide an adequate quantity of familiar-talker observations. The present design had used nineteen talkers to minimize potential talker-specific effects; a modified design could achieve the same desideratum by selecting four talkers who, in the present experiment, supported an equal proportion of illusions averaged across all untrained participants. Selecting these four talkers, and featuring them with equal frequency, should avoid an exposure effect among untrained listeners, and effects induced by training would be observed as deviations from these talkers' otherwise-equivalent integration levels.

Differences among high-integrator groups suffered from a general lack of power. This might be alleviated by providing only audiovisual training. Audiovisual training produced a stronger effect than did visual training, and auditory training produced no effect. Moreover, there was no suggestion from either high-integrator or low-integrator results that participants who received audiovisual training had responded differently than those who received visual training. Finally, untrained participants demonstrated an unambiguous bias for auditory percepts, whereas audiovisual participants were clearly biased toward visual responses. Therefore, audiovisual training should maximize differences between trained and untrained participants without overlooking differences that would have arisen from unimodal training.

2.3.5 Future directions

The present design raises the question of how listeners experience modal bias. Previously, listeners have been observed to be "more auditory" and "more visual" in their responses to audiovisual stimuli (Schwartz, 2010); however, in the present experiment, listeners were "more auditory" or "more visual" in how they learned to process a familiar talker, and not in how they responded to audiovisual stimuli. These results could support either of two interpretations of how biases may originate. Listeners may become biased toward one mode because they are less capable of processing the other; alternatively, biases may represent strategic differences arising from the quality of information obtained in prior learning. These alternatives are not exclusive of each other, and either could help to explain the disparity found in the present experiment. That is, listeners showed significant modal biases in how they learned and in how they responded, but these biases were not consistent with each other. This difference suggests that it may be too simplistic to claim that listeners "are" either auditory or visual. The present findings invite further study to determine under what conditions listeners will adopt or demonstrate modal biases.

2.4 Conclusion

The relative contributions of faces and voices to audiovisual speech integration seems to present a chicken-and-egg problem. Does a listener's integration skill cause him or her to attend more closely to one sensory channel, or does a listener's personal bias to attend to one sensory channel determine his or her integration skill? In the present

experiment, observers who performed efficient integration were more influenced by vision, and less-efficient observers appeared to prefer audition. However, all familiar-talker effects were driven by a familiar voice. In one view, this result could mean that "visual" listeners' focus on faces gave them more informational resources to successfully integrate, and that "auditory" listeners' disproportionate focus on voices destabilized integration. Conversely, it could be that "skilled" observers were more able to process the additional information presented in visual channels, whereas "unskilled" observers could not accommodate more than the auditory channel. These two perspectives essentially define the current discussion (Massaro & Cohen, 2000; Grant, 2002), and may be informed by the present findings. The present findings associate high integration with "visual listeners" and low integration with "auditory listeners," and demonstrate that modal bias in optimal integration may be affected by prior experience. Resolving the debate should provide us with greater understanding of the mechanism of optimal integration, and help us understand how different listeners can report different outcomes when perceiving audiovisual speech.

2.5 Acknowledgements

We would like to thank Katherine Jongsma, who assisted in collecting data, and the Natural Sciences and Engineering Research Council of Canada who supported D.I.S. through a Discovery Grant.

2.6 References

Abercrombie, D. (1964). Arthur the Rat. In *English Phonetics Texts* (pp. 117–119). London: Faber and Faber.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, *15*(9), 839–843. doi:10.1016/j.cub.2005.03.046

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research*, *183*(3), 399–404. doi:10.1007/s00221-007-1110-1

Aruffo, C., & Shore, D. I. (2012). Can you McGurk yourself? Self-face and self-voice in audiovisual speech. *Psychonomic Bulletin & Review*, *19*(1), 66–72. doi:10.3758/s13423-011-0176-8

Beauchemin, M., De Beaumont, L., Vannasing, P., Turcotte, A., Arcand, C., Belin, P., & Lassonde, M. (2006). Electrophysiological markers of voice familiarity. *European Journal of Neuroscience*, *23*(11), 3081–3086. doi:10.1111/j.1460-9568.2006.04856.x

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 445–463. doi:10.1037/0096-1523.30.3.445

Brédart, S., Barsics, C., & Hanley, R. (2009). Recalling semantic information about personally known faces and voices. *European Journal of Cognitive Psychology*, *21*(7), 1013–1021. doi:10.1080/09541440802591821

Barutchu, A., Crewther, S.G., Kiely, P., Murphy, M.J., and Crewther, D.P. (2008). When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology*, *20*(1), 1–11. doi:10.1080/09541440601125623

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 521–533. doi:10.1037/0278-7393.20.3.521

Colin, C., Radeau, M., & Deltenre, P. (2005). Top-down and bottom-up modulation of audiovisual integration in speech. *European Journal of Cognitive Psychology*, *17*(4), 541–560. doi:10.1080/09541440440000168

Craik, F. I. M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, *26*(2), 274–284. doi:10.1080/14640747408400413

Daßler, H., Gottschlich, K., Itz, M., Knösing, A., & Temmerman, M. (2008). Hörst Du mein Gesicht? - Audiovisuelle Integration. In *Programm 2008* (pp. 36–37).

De Gelder, B., Vroomen, J., Annen, L., Masthof, E., & Hodiamont, P. (2003). Audio-visual integration in schizophrenia. *Schizophrenia Research*, *59*(2–3), 211–218. doi:10.1016/S0920-9964(01)00344-9

Dodd, B. (1977). The role of vision in the perception of speech. *Perception*, *6*(1), 31–40. doi:10.1068/p060031

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends In Cognitive Sciences*, *8*(4), 162–169. doi:10.1016/j.tics.2004.02.002

Fairbanks, G. (1960). The Rainbow Passage. In *Voice and Articulation Drill Book* (pp. 124–139).

Grant, K. W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective. *Journal of the Acoustical Society of America*, *112*(1), 30–33. doi:10.1121/1.1482076

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America*, *104*(4), 2438–2450. doi:10.1121/1.423751

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America*, *103*(5), 2677–2690. doi:10.1121/1.422788

Green, K. P., Kuhl, P. K., & Meltzoff, A. N. (1988). Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment. *Journal of the Acoustical Society of America*, *84*(S1), S155. doi:10.1121/1.2025888

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, *50*(6), 524–536. doi:10.3758/BF03207536

Hampson, M., Guenther, F. H., Cohen, M. M., & Nieto–Castanon, A. (2003). Changes in the McGurk effect across phonetic contexts. In *CAS/CNS Technical Report Series* (p. 006).

Hardison, D. M. (1999). Bimodal speech perception by native and nonnative speakers of English: factors influencing the McGurk effect. *Language Learning*, *49*(s1), 213–283. doi:10.1111/0023-8333.49.s1.7

Hayashi, Y., & Sekiyama, K. (1998). Native-foreign language effect in the McGurk effect: A test with Chinese and Japanese. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*.

Hughes, S. M., & Nicholson, S. E. (2010). The processing of auditory and visual recognition of self-stimuli. *Consciousness and Cognition*, *19*(4), 1124–1134. doi:10.1016/j.concog.2010.03.001

Jones, J. a, & Jarick, M. (2006). Multisensory integration of speech signals: the relationship between space and time. *Experimental Brain Research*, *174*(3), 588–594. doi:10.1007/s00221-006-0634-0

Jones, J. A., & Munhall, K. G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics*, *25*(4), 13–19.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: matching identity across modality. *Current Biology*, *13*(19), 1709–1714. doi:10.1016/j.cub.2003.09.005

Kaufmann, J. M., & Schweinberger, S. R. (2005). Speaker variations influence speechreading speed for dynamic faces. *Journal of the Acoustical Society of America*, *34*(5), 595–610. doi:10.1068/p5104

Keyes, H., & Brady, N. (2010). Self-face recognition is characterized by "bilateral gain" and by faster, more accurate performance which persists when faces are inverted. *Quarterly Journal of Experimental Psychology*, *63*(5), 840–847. doi:10.1080/17470211003611264

Keyes, H., Brady, N., Reilly, R. B., & Foxe, J. J. (2010). My face or yours? Event-related potential correlates of self-face processing. *Brain and Cognition*, *72*(2), 244–254. doi:10.1016/j.bandc.2009.09.006

Lachs, L., & Pisoni, D. B. (2004a). Crossmodal source identification in speech perception. *Ecological Psychology*, *16*(3), 159–187. doi:10.1207/s15326969eco1603_1

Lachs, L., & Pisoni, D. B. (2004b). Specification of cross-modal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America*, *116*(1), 507–518. doi:10.1121/1.1757454

Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition*, *12*(3), 429–442. doi:10.1080/13506280444000382

Lander, K., & Davies, R. (2008). Does face familiarity influence speechreadability? *Quarterly Journal of Experimental Psychology*, *61*(7), 961–967. doi:10.1080/17470210801908476

Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). It's not what you say but the way you say it: matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 905–914. doi:10.1037/0096-1523.33.4.905

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, *26*(4), 708–715. doi:10.3758/BF03211391

MacDonald, J., Andersen, S., & Bachmann, T. (2000). Hearing by eye: how much spatial degradation can be tolerated? *Perception*, *29*(10), 1155–1168. doi:10.1068/p3020

Magnuson, J. S., Yamada, R. A., & Nusbaum, H. C. (1995). The effects of familiarity with a voice on speech perception. In *Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan* (pp. 391–392).

Massaro, D. W. (1987). *Perceiving talking faces: from speech perception to a behavioral principle*. (S. E. Palmer, Ed.) (p. 507). Cambridge, MA: MIT Press.

Massaro, D. W. (2002). From multisensory information to talking heads. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 153–176). Cambridge, MA: Bradford Books.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(5), 753–771. doi:10.1037/0096-1523.9.5.753

Massaro, D. W., & Cohen, M. M. (1990). Perception of synthesized audible and visible speech. *Psychological Science*, *1*(1), 55–63. doi:10.1111/j.1467-9280.1990.tb00068.x

Massaro, D. W., & Cohen, M. M. (2000). Tests of auditory–visual integration efficiency within the framework of the fuzzy logical model of perception. *Journal of the Acoustical Society of America*, *108*(2), 784–789. doi:10.1121/1.429611

Massaro, D. W., Cohen, M. M., Gesi, A., & Heredia, R. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics*, *21*(4), 445–478.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. doi:10.1038/264746a0

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*(1), 365–378. doi:10.1121/1.397688

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*(3), 351–362.

Munhall, K. G., ten Hove, M. W., Brammer, M., & Paré, M. (2009). Audiovisual integration of speech in a bistable illusion. *Current Biology*, *19*(9), 735–739. doi:10.1016/j.cub.2009.03.019

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustical Society of America*, *132*(2), 1061–1077. doi:10.1121/1.4728187

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., & Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Research*, *1323*, 84–93. doi:10.1016/j.brainres.2010.01.059

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. doi:10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. doi:10.1111/j.1467-9280.1994.tb00612.x

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328.

Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics*, *65*(4), 553–567. doi:10.3758/BF03194582

Putzar, L., Hötting, K., & Röder, B. (2010). Early visual deprivation affects the development of face recognition and of audio-visual speech perception. *Restorative Neurology and Neuroscience*, *28*(2), 251–257. doi:10.3233/RNN-2010-0526

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects. *Psychological Science*, *18*(5), 392–396. doi:10.1111/j.1467-9280.2007.01911.x

Rosenblum, L. D., Smith, N. M., Nichols, S. M., Hale, S., & Lee, J. (2006). Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception & Psychophysics*, *68*(1), 84–93. doi:10.3758/BF03193658

Rosenblum, L. D., & Yakel, D. A. (2001). The McGurk effect from single and mixed speaker stimuli. *Acoustics Research Letters Online*, *2*(2), 67–72. doi:10.1121/1.1366356

Rouger, J., Fraysse, B., Deguine, O., & Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Research*, *1188*, 87–99. doi:10.1016/j.brainres.2007.10.049

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences*, *104*(17), 7295–7300. doi:10.1073/pnas.0609419104

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception & Psychophysics*, *54*(3), 406–416. doi:10.3758/BF03205276

Sanchez, K., Dias, J. W., & Rosenblum, L. D. (2013). Experience with a talker can transfer across modalities to facilitate lipreading. *Attention, Perception & Psychophysics*, 1–7. doi:10.3758/s13414-013-0534-x

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology*, *65*(1), 111–116. doi:10.1037/0021-9010.65.1.111

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 915–930. doi:10.1037/0278-7393.18.5.915

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences*, *102*(51), 18748–18750. doi:10.1073/pnas.0508862102

Schwartz, J.-L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *Journal of the Acoustical Society of America*, *127*(3), 1584–1594. doi:10.1121/1.3293001

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research*, *40*(2), 453–463.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, *90*(4), 1797–1805. doi:10.1121/1.401660

Smith, E. G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry*, *48*(8), 813–821. doi:10.1111/j.1469-7610.2007.01766.x

Soto-Faraco, S., & Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *Neuroreport*, *18*(4), 347–350. doi:10.1097/WNR.0b013e32801776f9

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 580–587. doi:10.1037/a0013483

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215. doi:10.1121/1.1907309

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, *16*(3), 457–472. doi:10.1080/09541440340000268

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. doi:10.1016/j.neuropsychologia.2006.01.001

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision*, *8*(9), 1–11. doi:10.1167/8.9.14

Vatakis, A., & Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics*, *69*(5), 744–756. doi:10.3758/BF03193776

Von Kriegstein, K., Dogan, O., Grüter, M., Giraud, A.-L., Kell, C. A., Grüter, T., Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences*, *105*(18), 6747–6752. doi:10.1073/pnas.0710826105

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, *57*(8), 1124–1133. doi:10.3758/BF03208369

Yakel, D. A., Rosenblum, L. D., & Fortier, M. A. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics*, *62*(7), 1405–1412. doi:10.3758/BF03212142

Yarmey, A. D., Yarmey, A. L., & Yarmey, M. J. (1994). Face and voice identifications in showups and lineups. *Applied Cognitive Psychology*, *8*(5), 453–464. doi:10.1002/acp.2350080504

Chapter 3: What'd I say? Word-recognition accuracy for self-speech

3.1 Introduction

Do you understand your own voice better than others? Self-voice is a familiar voice (Kaplan, Aziz-Zadeh, Uddin, & Iacoboni, 2008), and familiar voices convey a speech-processing advantage (Nygaard, Sommers, & Pisoni, 1994). However, although we hear our own voices constantly, we hear self-voice through our head bones, in addition to through the air (Békésy, 1949). Bone-conducted speech does not sound the same as air-conducted speech, meaning that when we hear a recording of our own voice, our voice sounds differently to us than it does when we speak (Shuster & Durrant, 2003). In the modern era, we frequently hear our own recorded voice, and our accumulated exposure makes us experts at identifying the sound of our own recorded voice. Nonetheless, learning to recognize the sound of a voice does not necessarily mean learning to recognize what that voice is saying. We may not be experts at recognizing our own recorded speech, due to differences between spoken self-voice and recorded self-voice. The purpose of the present experiment was to determine whether recorded self-speech conveyed a familiar-talker advantage to word recognition.

Models of speech perception have more recently begun to address the role of familiar voices in speech processing. The issue is a natural outgrowth of the essential question of how listeners can understand different voices. Before the advent of electronic

analysis, acoustic differences among talkers were noted, but not regarded with great curiosity, as it was supposed that speech sounds were the natural consequence of common articulatory actions (Joos, 1948). Once the frequency components of speech could be electronically analyzed, however, it was revealed that different talkers produce substantially different acoustic realizations of the "same" speech actions (Peterson & Barney, 1952). No two talkers speak exactly alike because of physical differences between talkers and idiosyncrasies in speech production (e.g., Carr & Trill, 1964). Nonetheless, it was evident that listeners processed variability among talkers with apparent ease. New models of speech perception were formed around the question of how listeners resolve talker variability. An early answer was the *motor theory* (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), which posited that listeners perceived speech by mirroring a talker's movements in their own motor system and recognizing those movements as speech segments. Although the motor theory became well-known and widely cited, it did not gain many adherents, and the majority of its discussion has been arguments against it (Galantucci, Fowler, & Turvey, 2006)—such as the observation that non-speech stimuli, which have no gestural origin, can nonetheless be interpreted as speech sounds (Lane, 1965). Subsequent models of speech perception treated between-talker variability as an acoustical problem with a mathematical basis (for a review of models, see Luce & McLennan, 2005). These models were referred to as *abstract* models, as they supposed that speech comprised an inventory of invariant abstract forms, and that the process of speech perception was that of nullifying talker-specific acoustic features, or *normalizing* each voice, to uncover and extract the invariant

speech disguised within (for a review of normalization, see Johnson, 2005). Abstract models, however, assigned no function to a talker's unique vocal characteristics except as an obstacle to speech perception, and were therefore sometimes interpreted to imply that normalization stripped away and discarded talker-specific features (e.g., Halle, 1985). In reaction to this, an *exemplar* model was introduced (Goldinger, 1998). The exemplar model claimed that speech was dependent on individual talkers, such that listeners learned a language by storing an inventory of speech samples from different talkers in long-term memory. Novel speech could then be processed by comparing it to the various exemplars held in memory to find a best-fit match. An exemplar model, however, has not supplanted abstract models, because abstract models typically have made no claims about the role of individual talkers beyond those computations performed to normalize their variability (see Luce & McLennan, 2005). Any abstract model could, therefore, be conceptually modified to accommodate an influence of individual talkers on normalization. In any case, whether abstract or exemplar, current models of speech perception must now recognize how individual voices affect the efficiency of speech processing.

3.1.1 Familiar-talker speech advantages

Repeated exposure to an individual voice makes speech processing more efficient. Words presented in a previously-heard voice are recognized more quickly and more accurately (Craik & Kirsner, 1974; Palmeri, Goldinger, & Pisoni, 1993). Listeners appear to "adjust" to a known talker; words are recognized more quickly when spoken by a single talker, as opposed to switching among multiple talkers (cf. Creel & Bregman,

2011; Mullennix, Pisoni, & Martin, 1989). The processing efficiency conveyed by re-presented voices indicates that information about individual talkers is retained in short-term memory and is used to interpret subsequent speech featuring those talkers' voices. The information retained includes surface details about how speech was intoned, and at what pitch level (Church & Schacter, 1994; Schacter & Church, 1992). From exposure to a voice, listeners do not merely learn its general properties, but gain fine-grained knowledge of the speech segments it produces (Eisner & McQueen, 2005; Smith, 2007). The more a listener is exposed to a voice, the more its idiosyncratic speech production is impressed upon the listener's long-term memory, and the more "familiar" that voice becomes.

Familiar voices facilitate word recognition (Nygaard, Sommers, & Pisoni, 1994). This was observed by training listeners, over nine days, to identify ten different talkers by name. At the conclusion of training, listeners were tested for their ability to recognize novel words spoken both by these now-familiar talkers and by unfamiliar talkers. The task was made difficult by presentation in white noise, at four signal-to-noise ratios from 10 dB to –5 dB; and, at each ratio, listeners recognized words more accurately from familiar talkers. However, only listeners who successfully learned to identify the ten voices demonstrated this advantage. "Poor learners," who at the conclusion of training could not reliably identify the target voices, received no word-recognition advantage from "familiar" speech. Therefore, mere exposure is not sufficient to generate a familiar-talker advantage; rather, an advantage is dependent on a listener's ability to learn and remember vocal speech idiosyncrasies. We are constantly exposed to our own voices,

because we hear ourselves every time we speak, but we do not yet know whether ongoing exposure to our self-voice helps us learn and remember the vocal idiosyncrasies that will convey a familiar-talker advantage to the perception of recorded self-speech.

3.1.2 Recorded self-voice

Self-voices could not be tested in a general population until fairly recently. Tests of self-voice perception must present a listener with recordings of his or her own voice; and, because audio recordings are heard only through the air, they do not present the same vocal qualities as the bone-conducted speech we ordinarily hear when talking (Békésy, 1949). Consequently, before audio recording devices were commonly affordable, people who could identify their own recorded voice constituted a special population of radio announcers and vocal performers. Among this special population, recording artists could identify themselves with 100% accuracy, and public speakers 65% (Rousey & Holzman, 1967), whereas the accuracy of an ordinary population ranged between 38% and 55% (Holzman, Rousey, & Snyder, 1966; Olivos, 1967). Testing self-voice perception in the laboratory was therefore confounded, because listeners able to identify their own recorded voice were also trained to be explicitly aware of their own articulatory techniques, and this level of self-knowledge was not representative of a general population. However, a general population could not be tested, because most people rarely heard their own recorded voice.

Nowadays, most people frequently hear their own recorded voice. Recording devices are inexpensive and ubiquitous, and exposure to self-recordings begins at an early age. Children as young as age 4 can reliably identify their own voice

(Strömbergsson, 2009), and a general adult population can identify their own recorded voices with 90–95% accuracy (Hughes & Nicholson, 2010; Rosa, Lassonde, Pinard, Keenan, & Belin, 2008). We can therefore be confident that most adults will identify their own recorded voice at ceiling levels; and, consequently, a general population can now be tested to examine the effects of self-voice on speech processing.

3.1.3 Audiovisual self-speech

Self-voice conveys a processing advantage to audiovisual speech (Aruffo & Shore, 2012). This was demonstrated by presenting listeners with recordings of themselves, manipulated to produce the *McGurk effect* (McGurk & MacDonald, 1976). The McGurk effect is an audiovisual illusion in which an auditory syllable (e.g., /ba/) is dubbed onto a different visual syllable (e.g., /ga/); an observer, presented with these two conflicting percepts, integrates them into a third percept that is not physically present (e.g., /da/). The illusory effect is mandatory and automatic—listeners cannot avoid hearing the illusion, even when aware of its nature and under instructions to ignore it (Massaro & Cohen, 1983). However, the illusion is driven by *optimal integration* (cf. Ernst & Bülthoff, 2004): the principle that listeners will give "weight" to each channel according to its reliability, and arrive at a final integrated percept by combining inputs according to their assigned weight (Massaro, 2004). Therefore, the illusion may be manipulated by disrupting equivalence between the reliability of a talker's face and voice. When a face is perceived to be more reliable than a voice, optimal integration will produce a percept that is categorically visual (/ga/); conversely, a more-reliable voice can produce an auditory percept (/da/) (Massaro, 1987). Aruffo & Shore (2012) presented

listeners with illusory stimuli featuring the listeners' own faces and voices, either presented together or mismatched to unfamiliar voices and faces. Listeners experienced a weaker illusion from stimuli that presented both their face and voice together, as well as from stimuli that presented self-voice. Self-face produced no effect, but self-voice weakened the illusion whenever it was presented; it could therefore be concluded that self-voice had conveyed a speech-processing advantage to the auditory channel. However, an advantage realized from audiovisual McGurk stimuli does not necessarily predict an advantage for auditory word recognition. The illusion was presented as nonsense syllables, and listeners may respond differently to valid words (Grant & Seitz, 1998). Therefore, although we have evidence that self-voice facilitates audiovisual speech processing, we do not yet know whether self-voice may also convey a word-recognition advantage to auditory speech.

3.1.4 Self-speech recognition

Self-voice could convey a familiar-talker advantage to word recognition. Our brains respond to recorded self-voice as a familiar voice (Kaplan, Aziz-Zadeh, Uddin, & Iacoboni, 2008), and familiar voices convey an advantage to word recognition (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). The more familiar the voice, the greater the advantage it conveys (Magnuson, Yamada, & Nusbaum, 1995); that is, the magnitude of a familiar-talker advantage is proportional to the level of experience with that familiar voice. Family members' voices convey a greater advantage than newly-familiar talkers, who in turn convey a greater advantage than unfamiliar talkers.

Therefore, because adults have had years of exposure to their own voices, self-voice—as a highly-familiar voice—could convey a strong word-recognition advantage.

Exposure to one's own speaking voice, however, may not equate to experience with recorded self-speech. When we speak, the voice we hear is a combination of bone-conducted vibration and sound traveling through the air (Békésy, 1949), so that our spoken voice does not sound the same as its recorded playback (Shuster & Durrant, 2003). This means that speech segments will also sound differently when played back, and may not be recognized as familiar. We react to the sound of our own recorded voice even when we are not explicitly aware that it is our own (Holzman, Rousey, & Snyder, 1966; Douglas & Gibbins, 1983), but sometimes this reaction is overt revulsion and self-denial (Gur & Sackeim, 1979). Moreover, listening to recordings of our own voice may not help us become familiar with our own speech, because our brains may not "listen" properly. Certain neural processing centers involved in speech perception are deactivated when we hear our own recorded voice (Jardri et al., 2007), and our brains suppress attention to recordings of our own speech (Graux et al., 2012). This may be due to our brains' need to attend our own voice differently than other voices; otherwise, when we spoke, we might believe a disembodied voice was speaking to us instead (Green & Preston, 1981). When we do attend to recorded self-speech, our brains show special activation not present for other familiar voices, which could represent extra effort necessary for attending to and processing self-speech (Nakamura et al., 2001). Therefore, processing one's own recorded voice could be less efficient, which would result in a self-talker disadvantage.

3.1.5 The present design

To test whether recorded self-speech conveyed an advantage or disadvantage to word recognition, we used a procedure similar to that which initially demonstrated a familiar-talker advantage (Nygaard & Pisoni, 1994). Listeners transcribed monosyllabic words presented in white noise at four different signal-to-noise ratios. To avoid the obvious confound of participants being able to remember the words they had recorded, listeners recorded 800 different words, and did not return for testing until after two weeks had passed. Also, words were selected so that every word could be confused for at least one other among those recorded (e.g., *chess* and *chest*). To accommodate the possibility that word recognition could be affected by a "gender code" (Church & Schacter, 1994; see also Palmeri, Goldinger, & Pisoni, 1993) participants were presented at test with eight voices, four male and four female, one of which was their own. Because listeners react to their own voices irrespective of accurate identification (Douglas & Gibbins, 1983), identification accuracy was tracked by participants identifying the voices they heard (as "me" or "not me") in addition to transcribing words. Additional participants were recruited as controls for intelligibility; these participants did not hear their own voices, but were each assigned to a "self" voice and asked to identify that voice when it appeared. It was not expected that these participants would be successful in identifying an unfamiliar voice with which they had received no training, but the instruction ensured that all participants would be attending to identity as well as linguistic content. By comparing word-recognition accuracy between self-voice and other voices, as well as

between self-listeners and untrained listeners, we may determine whether self-voice exerts an effect on speech processing.

3.2 Experiment 1

3.2.1 Method

Participants

Fifty talkers (22 male, 28 female, age 17–24 years) were recruited as *self* participants. Forty-eight sex-matched controls (21 male, 27 female, age 18–23 years) were recruited as *surrogate* participants. All participants were native English-speaking McMaster University students who gave informed consent to the procedures and received course credit for their participation. All procedures complied with the tri-council ethics procedures in Canada as approved by the McMaster Research Ethics Board. Less course credit was available for each *surrogate* participant than for each *self* participant; each *surrogate* participant's testing session was therefore made half the length of a *self* participant's session.

Stimuli

Speech stimuli were recorded from 50 talkers. Talkers recorded 10 sets of 50 words each from phonetically-balanced lists (Egan, 1948) and 6 sets of 50 words each from the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965), for a total of 800 words. All participants read the words in the same order. Talkers were instructed to read at a natural pace, pausing between words. After completing the 800-word lists, talkers were asked to repeat any misspoken words (e.g, pronouncing *arch* as "arc" or

siege as "sedge"). Some words appeared on multiple lists, resulting in 710 unique words per participant. A grand total of 35,500 unique stimuli were created, forming a pool from which the experimental procedure drew.

When recorded, talkers were seated in a sound-attenuated room. Speech was recorded to WAV format (stereo, 16-bit depth, 44100 samples/s) with a Røde NT1000 condenser microphone and noise-reduced using Adobe Audition 3. Each word was individually normalized to a perceived loudness of –15 dB and converted to MP3 format (stereo, 128 kbps, 44100 samples/s).

A 180-minute white-noise sample, with a constant perceived loudness of –20 dB, was created using Adobe Audition 3 and saved in MP3 format (stereo, 128 kbps, 44100 samples/s).

Apparatus and procedure

Participants returned for testing no earlier than two weeks after recording. The testing session was programmed in Realbasic 2008 and presented on a 15" iMac G3 desktop computer running OSX. Audio was played through Maxell HP-200 stereo headphones with the computer's volume fixed at 80% of maximum.

The white-noise sample was played throughout the entire testing session. Before testing, participants were presented with the noise, at testing volume, and were solicited to decline participation if they felt that the noise would become intolerable. Participants were informed that, during the experiment, they could remove the headphones and take a break, or cease testing completely, with no negative consequence. Participants were

observed to take breaks during the procedure, but no participant declined to participate and no participant failed to complete a session.

Each trial required a listener to recognize a word and identify its talker. To begin each trial, an alerting asterisk appeared on screen for 500 ms, after which an auditory stimulus was presented. Participants typed the word, clicked one of two buttons labeled "me" or "not me," and then clicked a "next word" button to initiate the next trial. A progress bar indicated the proportion of trials completed.

Each *surrogate* participant was assigned to a sex-matched talker as a "self" voice. Each surrogate's session began with an on-screen instruction, "try to remember this person," followed by a single presentation of the word "are." This single exposure was not expected to induce a familiar-talker effect. During the testing session, surrogate participants selected "me" or "not me," but were not expected to successfully identify the "self" talker. Rather, this instruction was included to ensure that all participants were attending to talker identity. Surrogate participants received no feedback at any time as to whether their selections were correct.

Participants were cautioned that they would not be able to recognize every word, but that they should always type a guess rather than a non-answer (e.g., "didn't hear" or "I don't know"), and that typographical errors would be considered incorrect answers.

Each participant heard stimuli spoken by eight different talkers. Four male and four female voices were presented, one of whom was the self-voice (or "self"-voice). Talker selection was pseudorandomized so that all fifty talkers would be featured with equivalent frequency across all participants.

Stimuli were randomly selected at test according to word-list membership. For each of the eight talkers presented, one balanced-phoneme list (BP) and one onset–rime list (OR) was randomly selected, with the constraints that none of the ten BP lists would be selected twice and all six OR lists would be selected at least once. Because each list comprised 50 words, each *self* participant was thus presented with 100 unique words per talker. For *surrogate* participants, whose sessions were half the length of *self* participants, the first 25 words of each list were presented, for 50 unique words per talker.

Stimuli were presented at four different signal-to-noise ratios. Because each stimulus had been normalized to –15dB, and noise loudness was –20dB, stimuli were adjusted at test to +5, 0, –5, or –10dB, producing signal-to-noise ratios of 10, 5, 0, and –5dB, respectively.

Stimuli were not blocked, but presented in random order. Randomization was constrained so that no two consecutive stimuli presented the same word. *Self* participants were presented 200 stimuli at each of four loudness ratios (10, 5, 0, –5); each set of 200 stimuli comprised 8 talkers (4 male, 4 female) speaking 25 words each, for a total of 800 stimuli. Surrogate participants were presented with 100 stimuli at each loudness ratio, each set comprising 8 talkers speaking 12 or 13 words each, for a total of 400 words.

3.2.2 Results

Results were measured from 47 *self* participants (21 male, 26 female) and 48 sex-matched *surrogate* participants (21 male, 27 female). Although 50 self participants had been recorded, two self participants were excluded due to testing errors, and one self participant declined to return after controls had already been recruited.

Measures used

Word-recognition accuracy was measured as proportion of words correctly transcribed. Typographical errors were counted as incorrect answers. Homonyms were considered correct if no alternative pronunciation were possible (e.g., *rap* vs. *wrap*); however, if an apparent homonym could be disambiguated within a standard Ontario dialect (e.g., *hock* vs. *hawk*), the answer was considered incorrect.

Talker-identification accuracy was measured with the signal-detection formula for d', $Z_{hits} - Z_{false\ alarms}$ (Stanislaw & Todorov, 1999, p. 142). Because Z-scores cannot be calculated from raw means of 0 or 1, and participants had completed 800 trials, scores of 0 and 1 were adjusted by $\pm\frac{1}{1600}$. False alarms were calculated from same-sex trials only, because no participant mistook themselves for a member of the opposite sex.

Because processing speech and talker identity are similar processes, both types of judgment could arguably be described as either "recognition" or "identification" interchangeably. To maintain consistency, the present discussion referred to listeners' speech judgments exclusively as word *recognition*, and judgments of talker identity as talker *identification*.

Self-talker speech-processing advantage (*Word recognition)*

To determine whether *self* listeners demonstrated a self-talker speech-processing advantage, a repeated-measures mixed-design ANOVA (2 × 2 × 4 × 3) was performed on word-recognition data using between-subjects factors *self–surrogate* (self, surrogate) and *listener sex* (male, female) and within-subjects factors *noise level* (10 dB, 5 dB, 0 dB, –5 dB) and *talker identity* (same-sex, different-sex, self-talker). One main effect was

observed, for *noise level*, $F(3, 273) = 1070.90$, $p < .001$, indicating that word-recognition accuracy decreased with increasing noise. No main effects were observed for *self–surrogate*, $F(1, 91) = 0.99$, *ns*, for *listener sex*, $F(1, 91) = 0.03$, *ns*, or for *talker identity*, $F(2, 182) = 3.00$, *ns*. Two two-way interactions were observed, between *self–surrogate* and *talker identity*, $F(2, 182) = 3.92$, $p = .022$, and between *self–surrogate* and *noise level*, $F(3, 273) = 9.69$, $p < .001$. A three-way interaction was also observed among *listener sex*, *talker identity*, and *noise level*, $F(6, 273) = 3.08$, $p = .006$. The four-way interaction among all factors was not significant, $F(6, 273) = 1.36$, *ns*. The significant interactions were investigated further using t-tests.

An interaction between *self–surrogate* and *talker identity* indicated a potential effect of self-talker stimuli. Therefore, planned comparisons were performed on levels of *talker identity*, comparing between *self* and *surrogate* groups, using two-tailed independent-samples t-tests (Figure 3.1). These tests showed that *self* listeners recognized words from *self-talker* stimuli more accurately ($M = .42$, $SEM = .01$) than did *surrogate* listeners ($M = .37$, $SEM = .02$) from "self"-talker stimuli, $t(93) = 2.20$, $p = .03$. For *same-sex* talkers, accuracy was not different between *self* ($M = .40$, $SEM = .01$) and *surrogate* ($M = .40$, $SEM = .01$), $t(93) = 0.03$, *ns*; also, no difference was observed for *opposite-sex* talkers between *self* ($M = .38$, $SEM = .01$) and *surrogate* ($M = .37$, $SEM = .01$), $t(93) = 0.45$, *ns*. This between-groups result indicated a self-talker advantage for word recognition.

To discover whether the self-talker advantage had been driven by *self* stimuli, simple-effects tests were performed on *self* listeners and *surrogate* listeners for levels of

talker identity. Surrogate listeners showed no main effect of *talker identity*, $F(2, 94) = 1.90$, *ns*, whereas *self* listeners did show an effect, $F(2, 92) = 4.67$, $p = .012$. Planned comparisons were subsequently performed, using two-tailed paired-samples t-tests, on *self* listener word-recognition scores. *Self* listeners were more accurate when listening to *self-talker* stimuli ($M = .42$, $SEM = .01$) than to *opposite-sex* stimuli ($M = .38$, $SEM = .01$), $t(43) = 3.08$, $p = .004$, but showed no advantage for *self-talker* versus *same-sex* ($M = .40$, $SEM = .01$), $t(43) = 1.62$, *ns*. *Same-sex* and *opposite-sex* were also not different from each other, $t(43) = 1.40$, *ns*. These results indicate that *self* listeners showed an advantage for *self-talker* stimuli versus *opposite-sex* talkers, but not for *self-talker* versus *same-sex* talkers (Figure 3.1).

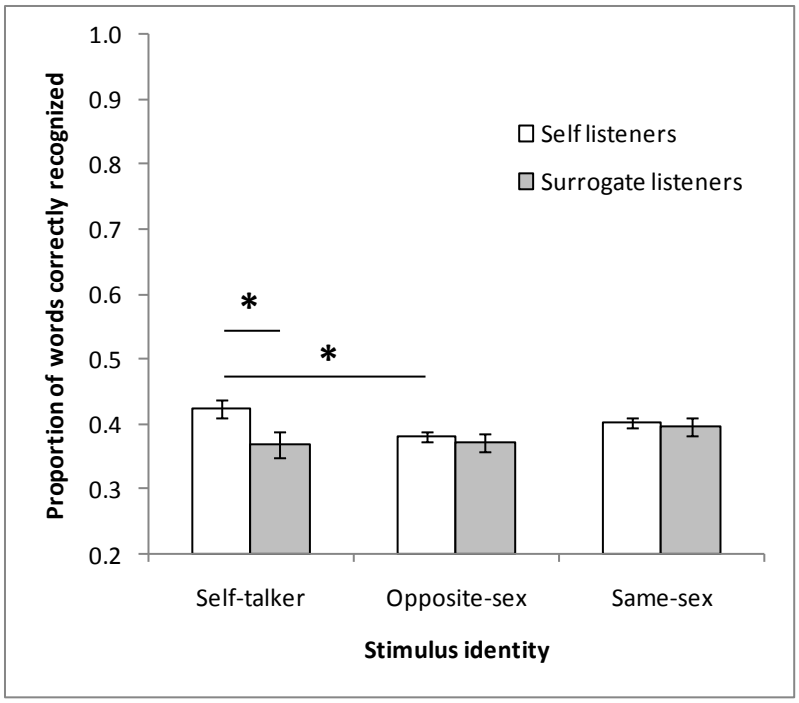

Figure 3.1 Word-recognition accuracy scores for *self* listeners and *surrogate* listeners. Error bars are standard error of the mean corrected for within-subjects design.

Two findings suggested that perception of self-talkers might change at different noise levels: the two-way interaction between *noise level* and *self–surrogate*, and the three-way interaction among *noise level*, *talker identity*, and *listener sex*. To examine the two-way interaction, independent-samples t-tests were performed on word-recognition accuracy, comparing between-group performance at each level of noise, corrected for multiple comparisons ($\alpha = .0125$). *Self* listeners demonstrated an overall advantage at 10 dB, the easiest listening condition, $t(93) = 3.03$, $p = .003$, and groups were not different at other noise levels. The three-way interaction was investigated with paired-samples t-tests, corrected for multiple comparisons ($\alpha = .0125$). These tests indicated that, regardless of whether listeners were *self* or *surrogate*, male listeners' word recognition was more accurate for *same-sex* stimuli at 0 dB and –5 dB, $t(47) = 3.59$, $p = .001$; $t(47) = 3.16$, $p = .003$. These interactions show, therefore, that a self-talker advantage was not affected by different noise levels; rather, men perceived male voices more accurately at lesser signal-to-noise ratios.

Self-identification and speech accuracy (d′ analysis)

To confirm the *a priori* expectation that *self* listeners would accurately identify their own voices, a repeated-measures ANOVA ($2 \times 2 \times 4$) was performed on voice-identification accuracy *d′* scores, using between-subjects factors *self–surrogate* (self, surrogate) and *listener sex* (male, female), and within-subjects factor *noise level* (10 dB, 5 dB, 0 dB, –5 dB). *Self* listeners were able to identify themselves, whereas *surrogate* listeners were unable to identify their assigned "selves," as shown by a main effect of *self–surrogate*, $F(1, 91) = 40.67$, $p < .001$. No effect was observed for *listener sex*, $F(1,$

91) = 0.42, *ns*. A two-way interaction between *self–surrogate* and *noise level*, $F(3, 273) = 5.29$, $p = .001$, resulted from *surrogate* participants being unable to identify their "self" voice in any noise condition, but *self* listeners' accuracy decreasing as noise ratios decreased: 10 dB, $M = 1.93$, $SEM = .21$; 5 dB, $M = 1.42$, $SEM = .14$; 0 dB, $M = 1.32$, $SEM = .15$; –5 dB, $M = 1.02$, $SEM = .12$ (Figure 3.2). *Self* listeners were, therefore, able to identify their own voice, with their accuracy depending on the level of noise.

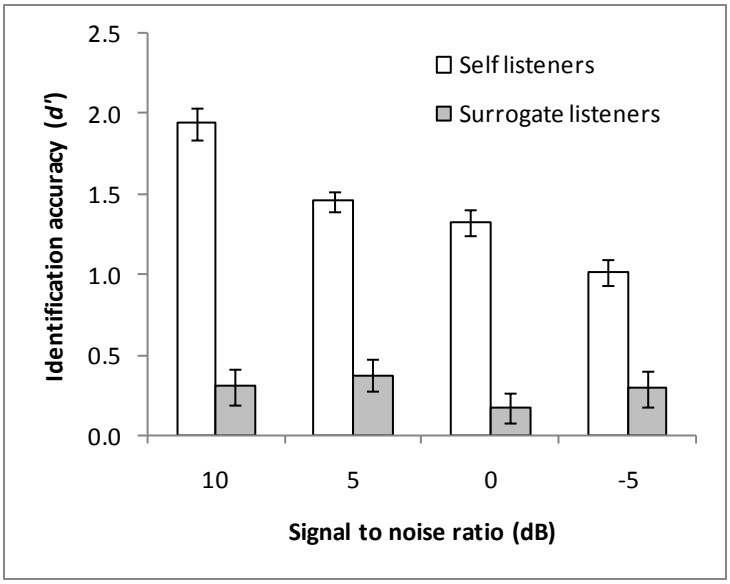

Figure 3.2 Identification accuracy at different noise levels. Error bars are standard error of the mean corrected for within-subjects design.

To examine whether self-identification was related to word-recognition accuracy, chi-square tests were performed. *Surrogate* listeners, who were unable to identify talkers, were excluded from these tests. Tests were performed on contingency tables constructed from individual trials, so as to determine on a trial-by-trial basis whether accurately identifying a talker influenced accurately recognizing a word. Contingency tables therefore comprised *talker identification* and *word recognition* for each trial. Separate tests were performed on *self* and *nonself* trials; and, because listeners did not mistake opposite-sex voices for their own, opposite-sex data were excluded from *nonself* trials. For *self* listeners in *self* trials, a relationship was observed between talker identification and word recognition, $X^2(1, N = 4448) = 126.56$, $p < .001$, $\varphi = .17$. For *nonself* trials, a relationship was also observed, $X^2(1, N = 13349) = 7.15$, $p = .007$, but the strength of this relationship was negligible ($\varphi = .02$).

Comparing expected and actual values for *self* listeners' *self* trials indicated the nature of the relationship between self-identification and word recognition (Table 3.1). When correctly identifying their own voice, listeners recognized words *more* accurately and made *fewer* errors. When failing to correctly identify their own voice, listeners recognized words *less* accurately and made *more* errors. These results show that word-recognition accuracy was affected by self-identification accuracy.

Table 3.1. Expected and actual values of word recognition for correct and incorrect talker identification, self-voice trials only. Values have been transformed to proportion of total trials.

		Expected	Actual
Correctly identified self	Word correct	.26	.30
	Word incorrect	.37	.33
Failed to identify self	Word correct	.15	.11
	Word incorrect	.21	.25

3.2.3 Discussion

Listeners transcribed words presented in noise, spoken by themselves and others, and identified whether the talker's voice was their own. Surrogate listeners were assigned to "self" voices and performed the same task. Two main results were found. Between groups, *self* listeners performed word recognition more accurately for their own speech than did *surrogate* listeners, confirming a self-talker advantage for word recognition. Within groups, *self* listeners recognized their own speech more accurately than opposite-sex speech, but not more accurately than same-sex speech. A post-hoc analysis on talker identification indicated that a self-talker advantage was dependent on accurate self-identification; however, no significant relationship was found between identifying nonself talkers as "not me" and word-recognition accuracy for nonself voices. These results therefore confirmed a self-talker advantage for word recognition, but could not fully disambiguate self-voice from same-sex voices.

A self-talker advantage was observed between subjects. That is, when presented with self-voice, *self* listeners recognized words more accurately than did *surrogate* listeners assigned to those same voices. This finding supports the hypothesis that self-voice, as a familiar voice, conveys a familiar-talker advantage.

Self listeners did not show an advantage for self-speech versus same-sex voices. Additionally, same-sex word-recognition accuracy was not different across participants; that is, *self* listeners and *surrogate* listeners were equally accurate for same-sex voices. This could be explained by the fact that listeners were actively listening for a target voice of a particular sex. Listeners can facilitate speech processing by adjusting to a

"coordinate system" relative to the voices they are hearing (Johnson, 2005). The present procedure featured voices of both sexes; listeners trying to identify a voice could have taken opposite-sex voices as a cue to attend more carefully to same-sex voices, and subsequent adjustment to sex-typical characteristics would have facilitated speech processing for all voices of that sex, conveying a same-sex advantage to all listeners.

In noise, male listeners perceived male voices more accurately at 0 and –5 dB than they did at 5 or 10 dB. This may be due to the fact that low-frequency sounds are less susceptible to environmental attenuation (Eyring, 1946), from which male voices could be easier to hear in noisier conditions. Female listeners did not show a similar advantage for male voices in noise, which could be taken as further support for listeners being oriented to a same-sex frame of reference.

Identification of talker sex remained at ceiling in all noise conditions, evidenced by the fact that listeners did not mistake opposite-sex voices to be self-voice. That is, if noise had made it difficult to judge sex characteristics of a voice, listeners would have demonstrated at least some confusion between self-voice and opposite-sex voices, and this did not occur. However, increased noise made explicit identification of self-voice more difficult. Because noise differently affected judgments of talker sex and explicit talker identity, it may be that *self* listeners found it easier to orient to the more-perceptible vocal qualities of same-sex voices than to attempt to discern their own voice's unique qualities. This could help explain why a self-talker advantage was not greater than a same-sex advantage. If *self* listeners were attending primarily to sex-typical vocal features, but nonetheless received an advantage from the presence of self-voice

characteristics, then the advantage conveyed by self-voice may have been sufficient to produce a difference in accuracy versus surrogates, but not sufficient to supersede the advantage conveyed by a same-sex frame of reference.

Word-recognition accuracy was nonetheless affected by self-identification accuracy. When *self* listeners correctly identified their own voices, they recognized words more accurately; when they failed to correctly identify their own voices, they recognized words less accurately. This significant effect was not due to stimuli being generally easier or more difficult to perceive, because a relationship was not observed between correctly identifying nonself voices (as "not me") and more-accurate word recognition.

The results of Experiment 1 show a speech-processing advantage for self-voice, and further show that this advantage is related to correctly identifying one's own voice. However, a self-talker advantage could not be fully disambiguated from an apparent same-sex advantage, and the relationship between correct talker identification and accurate word recognition could not be fully explained. Experiment 2 was designed to address these issues.

3.3 Experiment 2

Experiment 2 was designed to disambiguate self-talkers from same-sex talkers and to further explore the relationship between correct self-identification and increased word-recognition accuracy.

To disambiguate self-talkers from same-sex talkers, Experiment 2 presented stimuli in noise and in the clear, and presented only same-sex talkers. Noise in Experiment 1 may have obscured talkers' individual vocal qualities, and the presence of opposite-sex voices may have cued listeners to rely instead on sex-typical vocal characteristics. Presenting stimuli in the clear should allow listeners to attend to self-voice qualities; however, presenting words in the clear runs the risk of participants recognizing words with ceiling accuracy. Experiment 2 therefore also featured one noise level. Because male voices had been better perceived at 0 dB and –5 dB signal-to-noise ratios, Experiment 2 presented words in noise at a ratio of 2.5 dB. If listeners gain a familiar-talker advantage from being able to hear the unique qualities of their own voice, and noise obscures those qualities, then self-speech in the clear should show greater word-recognition accuracy than same-sex voices in the clear. If, however, listeners do not gain an advantage from self-voice qualities, but can only adapt to a same-sex frame of reference, then same-sex and self-voice would show equivalent accuracy, with or without noise.

Experiment 1 demonstrated a relationship between correctly identifying a talker and accurately recognizing a word, but could not suggest whether either one caused the other. Experiment 2 examined whether word recognition had contributed to talker identification. Two manipulations were introduced: temporally-reversed stimuli, and a participant group that did not transcribe words. If transcribing words caused participants to notice talker-specific characteristics that facilitated talker identification, then transcribing participants should identify talkers more accurately than non-transcribing

participants. Reversed speech should help disambiguate an effect of speech content versus an effect of attention. Reversed speech preserves many of a talker's vocal qualities, but eliminates all valid speech information; consequently, a talker can be identified from reversed speech, although less accurately (Sheffert, Pisoni, Fellowes, & Remez, 2002; van Lancker, Krieman, & Emmorey, 1985;). If correct identification was facilitated by greater attention to each stimulus, then transcribing reversed words should oblige listeners to pay more attention, and talkers would be identified from reversed words more accurately by transcribing participants than by non-transcribers. By examining the effect of transcription, for normal and temporally-reversed stimuli, Experiment 2 explored whether familiar speech influenced talker identification.

3.2.1 Method

Participants

Forty-one *self* participants (7 male, 34 female, age 18–28) and 82 sex-matched *surrogate* participants (14 male, 68 female, age 17–24), all native English speakers, were newly recruited and gave informed consent to the procedure. Participants were McMaster University students and were compensated with course credit. None of the participants from Experiment 1 participated in Experiment 2. As with experiment 1, *surrogate* participants' testing session was half the length of *self* participants; however, in Experiment 2, this was balanced by recruiting twice as many *surrogate* participants.

Stimuli

Stimuli were recorded from 41 *self* participants under the same conditions as Experiment 1. All new talkers recorded the same 800 words as the previous experiment,

but only 100 words were used from each talker for the present procedure (the same two 50-word balanced-phoneme lists). Each stimulus was also temporally reversed using Adobe Audition 3; onset and offset of each stimulus were identified as a complete cessation of the signal at either end. A total of 8,200 new stimuli were thereby created for Experiment 2.

Six nonself talkers, 3 male and 3 female, were used. These six talkers were selected as the most intelligible talkers from Experiment 1 averaged across all conditions. These talkers' stimuli were temporally reversed, resulting in 1,420 total stimuli for each talker and 8,520 stimuli total. A grand total of 16,720 stimuli were thus available to the current experiment, in addition to the same white-noise sample used in Experiment 1.

Apparatus and procedure

Apparatus and instructions were identical to Experiment 1. The procedure was identical except for the following differences.

Self participants were randomly divided into *transcription* and *no-transcription* groups. One stimulus was presented in each trial; after its presentation, *transcription* participants transcribed the word and *no-transcription* participants did not. Two *surrogate* participants were pseudorandomly assigned to each "self" talker and to the same *transcription* or *no-transcription* group as the self participant.

Stimuli were presented in two noise conditions. *Clear* trials presented words at a perceived loudness of –15 dB. Because word-recognition accuracy was greater for female voices at 10 or 5 dB loudness ratios, and greater for male voices at 0 or –5 dB,

noisy trials featured white noise at –20 dB, and stimuli at –17.5 dB, for a signal-to-noise ratio of 2.5 dB.

Stimuli were organized into four blocks: *clear–forward*, *clear–reverse*, *noisy–forward*, and *noisy–reverse*. Block order was pseudorandomized to provide equivalent frequency for each possible ordering, with the constraint that both *noisy* blocks were always consecutive. For self participants, each block comprised four lists of 50 words, spoken by the four talkers, for a total of 200 words in each block and a grand total of 800 trials. Surrogate participants listened to 100 words in each block for a grand total of 400 trials. Word order was randomized with the constraint that two consecutive stimuli did not present the same word or the same talker. Only balanced-phoneme word lists were used. Only same-sex nonself talkers were presented to each participant.

3.2.2 Results

Results were measured from all 41 *self* participants and 82 *surrogate* participants. Surrogates were sex-matched. One surrogate female was inadvertently presented with male voices; these data were included, as the same analyses were run with and without these data and results were not different. Word-recognition accuracy and talker-identification accuracy were measured as in Experiment 1. Word-recognition scores were not recorded for temporally-reversed stimuli or for participants who did not transcribe words.

All ANOVA tests described below were initially performed inclusive of the factor *talker sex*. This factor produced no effect or interaction in any test, and has been excluded for the sake of clarity.

Self-talker speech-processing advantage

To assess whether *self* listeners demonstrated a self-talker speech-processing advantage, a mixed-design repeated-measures ANOVA (2 × 4 × 3) was performed on word-recognition data using between-subjects factors *self–surrogate* (self, surrogate) and within-subjects factors *noise level* (noise, clear) and *talker identity* (nonself, self-talker). Data from *reversed-direction* trials and from *non-transcriber* participants were excluded, as they presented no valid measures of word-recognition accuracy. *Self* listeners were more accurate overall, indicated by a main effect of *self–surrogate*, $F(1, 62) = 16.05$, $p < .001$. Word-recognition accuracy was greater in the clear than in noise, shown by a main effect of *noise level*, $F(1, 62) = 722.84$, $p < .001$. *Noise level* interacted with *self–surrogate*, $F(1, 62) = 4.36$, $p = .041$, and a three-way interaction was observed among *noise level, self–surrogate*, and *talker identity*, $F(1, 62) = 19.44$, $p < .001$.To interpret these interactions, planned comparisons were performed between *self* listeners and *surrogate* listeners, and between *self-talker* and *nonself* stimuli, in each noise condition.

Word-recognition accuracies for *self-talker* and *nonself* stimuli were compared between-subjects and within-subjects for each level of noise (Figure 3.3). Between-subjects comparisons were performed with independent-samples t-tests, and within-subjects comparisons were performed with paired-samples t-tests. Between-subjects comparisons showed that *self* listeners recognized *self-talker* stimuli more accurately than did *surrogate* listeners, both in the clear, $t(62) = 3.27$, $p = .002$, and in noise, $t(62) = 3.26$, $p = .012$. Between-subjects accuracy for *nonself* stimuli was not different in the clear, $t(62) = 0.68$, *ns*, but *self* listeners were more accurate for *nonself* stimuli in noise, $t(62) =$

2.60, $p = .012$. Within-subjects comparisons between *self-talker* and *nonself* stimuli showed that, in the clear, *self* listeners recognized *self-talker* stimuli more accurately, $t(20) = 2.38$, $p = .028$, whereas *surrogate* listeners recognized *self-talker* stimuli less accurately, $t(42) = –4.08$, $p < .001$. In noise, no within-subjects difference was observed either for *self* listeners, $t(20) = –1.00$, or for *surrogate* listeners, $t(42) = 1.06$, *ns*. These results indicated that *self* listeners did show a self-talker advantage and, in noisy conditions, received an equivalent advantage from same-sex nonself talkers.

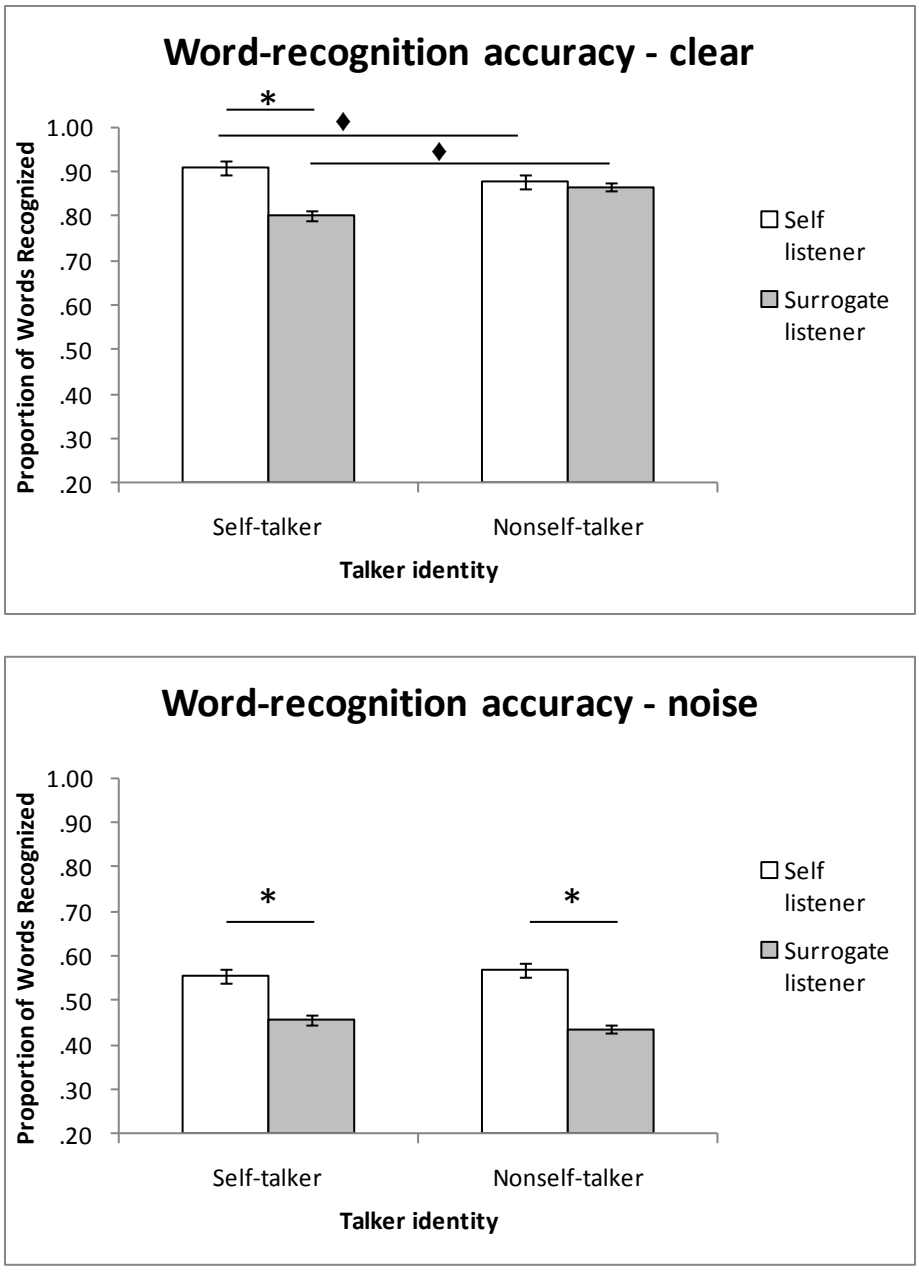

Figure 3.3 Word-recognition for *self* and *surrogate* listeners, showing *self-stimuli* and *nonself* stimuli across noise conditions. Between-groups differences are marked with asterisks; within-group differences are marked with diamonds. Error bars are standard error of the mean corrected for within-subjects design.

Self-identification and speech accuracy

To examine whether performing the transcription task affected listeners' ability to identify talkers, identification-accuracy scores were tested. If there were an effect, it would be observed as differences between *transcriber* and *non-transcriber* groups, shown for either *normal-direction* or *reversed-direction* stimuli. A mixed-design repeated-measures ANOVA was therefore performed on identification-accuracy data using between-subjects factors *self–surrogate* (self, surrogate) and *transcriber* (transcriber, non-transcriber) and within-subjects factors *noise level* (clear, noise) and *stimulus direction* (normal, reversed). The factor of interest, *transcriber*, produced no effects or interactions, indicating that attention to word transcription did not affect talker identification in any condition.

To examine whether self-identification was related to word-recognition accuracy, chi-square tests were performed as with Experiment 1. That is, tests were performed on contingency tables comprising *talker identification* and *word recognition*, for both *self* and *nonself* trials, using individual-trial data. *Surrogate* listeners, being unable to identify talkers, were excluded. Because no word-recognition advantage had been observed for self-voice in noise, separate tables were tested for *noisy* and *clear* trials. These relationships were significant for self trials in the clear, $X^2(1, N = 1050) = 16.02$, $p < .001$, $\varphi = .12$, and in noise, $X^2(1, N = 1050) = 14.64$, $p < .001$, $\varphi = .12$. Because word-recognition accuracy in noise was equivalent for both self-voice and nonself-voice, nonself data were also tested for an association between correct identification (correctly rejecting a *nonself* voice) and accurate word recognition. These tests showed no

significance either in the clear, $X^2(1, N = 3150) = 3.11$, *ns*, or in noise, $X^2(1, N = 3150) = 0.18$, *ns*.

Comparing expected and actual values of *self* trials, excluding *nonself* trials, indicated the relationship between talker identification and word recognition. All values were converted to a proportion of total trials for easier comparison. Both in the clear and in noise, the effects were the same (Table 3.2): listeners recognized words more accurately and made fewer errors when correctly identifying their own voice, and recognized words less accurately and made more errors when failing to identify their own voice. These results show that word-recognition accuracy was affected by self-identification accuracy.

Table 3.2. Expected and actual values of word recognition for correct and incorrect talker identification, self-voice trials only. Values have been transformed to proportion of total trials.

Clear

		Expected	Actual
Correctly identified self	Word correct	.77	.78
	Word incorrect	.08	.06
Failed to identify self	Word correct	.14	.13
	Word incorrect	.01	.03

Noise

		Expected	Actual
Correctly identified self	Word correct	.41	.43
	Word incorrect	.33	.30
Failed to identify self	Word correct	.15	.12
	Word incorrect	.12	.15

3.2.3 Discussion

Listeners transcribed words presented in noise and in the clear, presented normally and temporally reversed, spoken by themselves and other nonself voices, and identified whether the talker's voice was their own. In the clear, *self* listeners recognized words more accurately from *self-talker* stimuli than from *nonself* stimuli, whereas *surrogate* listeners recognized words less accurately from *"self"-talker* stimuli than from *nonself* stimuli. In noise, *self* listeners recognized all words more accurately than *surrogate* listeners; accuracy did not differ between *self-talker* and *nonself* stimuli. Transcribing words had no effect on talker identification. A post-hoc evaluation indicated that. These results indicated that self-speech had conveyed a self-talker advantage to word recognition in the clear, and that the presence of self-speech had conveyed a general advantage to word recognition in noise.

In the clear, a self-talker advantage for word recognition was evident. *Self* listeners recognized self-speech *more* accurately than speech from other talkers, whereas *surrogate* listeners recognized "self"-speech *less* accurately than speech from other talkers. It is not surprising that *surrogate* listeners recognized "self" speech less accurately, because the other talkers were preselected for their high intelligibility, and "self" talkers had been newly recruited. However, this result underscores *self* listeners' superior accuracy in recognizing their own speech. A self-talker advantage obtained only in the clear, and not in noise, indicating that noise had obscured whatever qualities of self-voice conveyed a self-talker advantage.

However, in noise, *self* listeners demonstrated a general word-recognition advantage versus *surrogate* listeners. In noise, *self* listeners found self-voices and nonself-voices equivalent in their intelligibility, but showed a generally-higher level of accuracy than did *surrogate* listeners. This suggests that, because no opposite-sex listeners had been presented, *surrogate* listeners were not cued to orient to sex-typical qualities, and therefore did not gain an advantage from attending to same-sex voices. By contrast, when *self* listeners could not identify the unique qualities of their own voice, they instead oriented themselves to its sex-typical qualities, which facilitated speech processing for all stimuli.

Self-identification accuracy was nonetheless related to word-recognition accuracy, both in noise and in the clear. Self-identification accuracy was not affected by attention to stimulus content, as shown by the act of transcribing words having no effect, even when the stimuli were reversed and thus difficult to process. This result suggests, instead, that correctly identifying one's own voice contributes to a self-talker advantage, presumably by prompting a listener to orient to the reference frame of his or her own vocal qualities. That this relationship should appear in noise, even when *self* listeners did not demonstrate a recognition advantage for self-voice, may be interpreted as the orientation to self-vocal qualities having a significant influence, but nonetheless less of an influence than a general orientation to sex-typical characteristics.

3.3 General discussion

The present investigation is the first to have tested self-voice perception for a familiar-talker advantage in word-recognition accuracy. In two experiments, self-voice, same-sex voices, and opposite-sex voices were presented at different signal-to-noise ratios. Listeners transcribed words and identified whether the voices were their own. When no noise was present, self-voice conveyed a familiar-talker advantage. When stimuli were presented in noise, same-sex voices conveyed an advantage equivalent to self-voice; nonetheless, in all conditions, correctly identifying self-voice contributed to a self-talker advantage for word recognition, indicating that a self-talker advantage was dependent on perceiving one's own familiar vocal qualities.

The present findings therefore support an interdependence of talker identity and speech processing, consistent with a familiar-talker advantage (Nygaard, Sommers, & Pisoni, 1994), such that increased perceptual learning of talker-specific vocal features also increases perceptual sensitivity for linguistic content. The present findings suggest, however, that perceptual learning need not be talker-specific. Previous tests of familiar-talker advantage had concluded that becoming able to identify a talker's voice was a "necessary but not sufficient condition" for showing a benefit to word recognition from that voice, implying that a speech-processing advantage was gained only from attention to the acoustic details that uniquely distinguished a familiar talker's voice from others (Nygaard & Pisoni, 1998). The current investigation demonstrated that, under conditions where a familiar talker's voice could not be readily identified, orientation toward a familiar talker could also facilitate speech from similar-sounding unfamiliar voices.

3.3.1 Response to self-voice

Hearing one's own voice did not suppress word recognition. It is important for us to distinguish between our own speech production and that of others, so that we do not "hear voices" when we talk (Green & Preston, 1981). However, the current findings provide evidence that the neural mechanisms that suppress speech-processing activity while producing self-voice, and which suppress attentional orientation to self-voice (Jardri et al., 2007; Graux et al., 2012), do not prevent us from attending to our own speech and developing a familiar knowledge of how we talk. Therefore, when we hear a recorded talker whom we identify as ourselves, we can bring our perceptual learning to bear upon that speech. A conflict is introduced when we identify recorded self-speech as our own voice, because we must also reject what we hear as being something we are not currently saying. The additional brain activity observed when processing self-speech may not be due to speech-processing difficulty, as suggested by Nakamura et al. (2001), but could be a consequence of listeners' need to correctly attribute the source of their own recorded speech production.

Identifying one's own voice was not shown to contribute to a self-talker advantage. Although Experiment 1 showed a potential association between correct identification of self-voice and increased word-recognition accuracy, Experiment 2 failed to find a similar association. This result could mean that the design of Experiment 2 experiment was insufficient to detect a relationship between these two measures. Alternatively, it could mean that a self-talker advantage for speech processing, under the conditions of Experiment 2, is independent of identifying one's own voice. Experiment 2 also showed no effect of transcribing words on talker identification; this might be taken

as support for the independence of familiar-speech processing and familiar-talker identification. These results, showing no effects of talker identification or word recognition on each other, are not conclusive, and thus bear further investigation.

3.3.2 Same-sex advantage

Attention to same-sex qualities conveyed a same-sex advantage, but only when listeners were actively listening for those qualities. When self-voice could be clearly identified, and was therefore more diagnostic than talker sex, no advantage was conferred to same-sex voices. By contrast, when stimuli were presented in noise, *self* listeners demonstrated an overall processing advantage for voices of the same sex, including their own voice. This result echoes the result found when testing self-talkers in audiovisual speech (Aruffo & Shore, 2012); in that experiment, the presence of self-speech induced a general processing advantage for all talkers, as well as a particular advantage to self-speech. The audiovisual-speech result for nonself voices can be explained without reference to self-speech qualities; because an advantage was conferred by self-voice, and not self-face, listeners could have been conditioned to pay greater attention to the auditory channel generally, which would affect integration of all stimuli. In the present experiment, only auditory speech was presented, meaning that the effect conveyed to same-sex voices could not be due to sensory-channel bias but may be attributable to familiar vocal qualities.

3.3.3 Speech-processing models

The present investigation into familiar speech was not designed to test speech-processing models directly, but its findings may contribute to further understanding of

how talker identity may be represented in speech processing. The present results may be interpreted either from an abstract or an exemplar perspective.

An abstract model of speech processing claims that talker identity is stored separately from abstract speech forms. Therefore, when a familiar talker's speech is identified, perceptual memory of that talker's voice may be recalled and used as a frame of reference to achieve more-efficient normalization of speech (cf. Johnson, 2005). In this model, the present findings demonstrated that when a talker's speech was identified as similar to a familiar talker, perceptual memory of its familiar features were recalled and used to facilitate speech normalization. This result supports granular encoding of talker identity; that is, a talker's voice may be encoded in different levels of detail, from general to idiosyncratic (Baumann & Belin, 2010). Vocal identities may therefore be encoded relative to a prototype (cf. Belin, Bestelmeyer, Latinus, & Watson, 2011), such that certain vocal qualities are identified as a set of shared prototypical characteristics, and individual voices within that set are further defined by their idiosyncratic deviations from the prototypical basis. The level of talker-identity detail recalled for use in speech normalization would, therefore, depend on the level of specificity to which it were possible to identify a voice. An abstract model could thus explain the results observed in the present experiment to be listeners achieving speech normalization by recalling different levels of familiar detail, in accordance with the level at which a talker could be identified.

An exemplar model of speech processing claims that talker-identifying characteristics are stored as elements of speech (Goldinger, 1998). To process speech,

our minds retain an inventory of previously-heard speech tokens, each of which represents an instance of that token as produced by a particular talker, and we interpret new speech through best-fit comparisons to old tokens (cf. Tenpenny, 1995). A familiar-talker advantage could therefore be explained by the efficiency of knowing which stored tokens should be retrieved; such a mechanism would explain the present result shown for self-speech in the clear. The same-sex advantage observed in noise suggests a similar efficiency, except that instead of talker identity indicating which specific token to retrieve, identifying a particular class of talker (e.g., "female") would indicate a particular class of token to retrieve. In other words, processing efficiencies in an exemplar model could be achieved not only by retrieving talker-based exemplars but by retrieving exemplars known to be in a similar class to that talker. If this were so, the present results could have implications for an exemplar model's description of how we organize speech tokens in memory. An exemplar model could explain the speech-processing efficiencies demonstrated in the present experiment as efficiencies of retrieval: the presentation of a familiar talker activated the retrieval of a matching token, or class of tokens, enabling a more-efficient search for the appropriate target.

3.3.4 Future directions: Learning mechanisms of self-voice

Whether we learn vocal qualities as prototypes or bundled with speech tokens, we do not yet know how we learn the qualities of self-voice that facilitate our perception of recorded self-speech. We cannot identify our own recorded voices without having had substantial exposure to such recordings (Holzman, Rousey, & Snyder, 1966), because our bone-conducted "internal" voice sounds differently than our air-conducted "external"

voice (Békésy, 1949). Furthermore, even though we now have substantial exposure to self-recordings, and can easily identify our own voices (e.g., Hughes & Nicholson, 2010), we do not know whether this exposure is what makes our own speech familiar to us. Becoming able to identify a recorded voice is not necessarily sufficient to induce a familiar-talker advantage for that voice (Nygaard & Pisoni, 1998), so it cannot be assumed that the exposure that trains us to identify our recorded voice is sufficient to induce a self-talker advantage. Moreover, speech patterns can themselves be made familiar and identifiable, independently of vocal quality (Fellowes, Remez, & Rubin, 1997). It is possible that we become familiar with our own speech patterns by hearing our own "internal" speech as we talk, and identifying our recorded voice merely facilitates our activating that speech-pattern familiarity when processing recorded self-speech. It might be possible to test this distinction by presenting listeners with self-speech whose quality has been adjusted to sound less like self-speech, but whose articulatory manner has been preserved. If a self-talker advantage were due to learning "internal" speech patterns, and not dependent on vocal quality, then listeners should demonstrate an advantage for self-speech even when the quality of the voice is unidentifiable. A pilot version of this proposed experiment was attempted using the voice-morphing algorithm STRAIGHT (Kawahara & Irino, 2005), but stimulus intelligibility was significantly deteriorated by the morphing process. More-current algorithms may be more successful at morphing without reducing intelligibility. Determining our level of awareness of our own articulatory habits could help us better

understand how we learn to pronounce words and, thereby, offer insights to articulatory-training pedagogy.

3.4 Conclusion

Self-voice conveys a self-talker advantage to speech perception. This advantage is dependent on being able to positively identify the voice. When a voice is obscured by noise and cannot be identified, an advantage is conveyed to all voices that generally resemble the target voice. We conclude from these findings that listeners successfully learn to perceive recorded self-speech as familiar speech, despite its sounding differently from bone-conducted speech.

3.5 Acknowledgements

We would like to thank Swapna Krishnamoorthy, who assisted in collecting data, and the Natural Sciences and Engineering Research Council of Canada who supported D.I.S. through a Discovery Grant.

3.6 References

Aruffo, C., & Shore, D. I. (2012). Can you McGurk yourself? Self-face and self-voice in audiovisual speech. *Psychonomic Bulletin & Review*, *19*(1), 66–72. doi:10.3758/s13423-011-0176-8

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110–120. doi:10.1007/s00426-008-0185-z

Békésy, G. V. (1949). The structure of the middle ear and the hearing of one's own voice by bone conduction. *Journal of the Acoustical Society of America*, *21*(3), 217–232. doi:10.1121/1.1906501

Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review*, *2*(3), 339–363. doi:10.3758/BF03210972

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 711–725. doi:10.1111/j.2044-8295.2011.02041.x

Carr, P. B., & Trill, D. (1964). Long-term larynx-excitation spectra. *Journal of the Acoustical Society of America*, *36*(11), 2033–2040. doi:10.1121/1.1919319

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 521–533. doi:10.1037/0278-7393.20.3.521

Craik, F. I. M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, *26*(2), 274–284. doi:10.1080/14640747408400413

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass*, *5*(5), 190–204. doi:10.1111/j.1749-818X.2011.00276.x

Douglas, W., & Gibbins, K. (1983). Inadequacy of voice recognition as a demonstration of self-deception. *Journal of Personality and Social Psychology*, *44*(3), 589–592. doi:10.1037/0022-3514.44.3.589

Egan, J. P. (1948). Articulation testing methods. *The Laryngoscope*, *58*(9), 955–991. doi:10.1288/00005537-194809000-00002

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. doi:10.3758/BF03206487

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends In Cognitive Sciences*, *8*(4), 162–169. doi:10.1016/j.tics.2004.02.002

Eyring, C. F. (1946). Jungle acoustics. *Journal of the Acoustical Society of America*, *18*(2), 257–270. doi:10.1121/1.1916362

Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics*, *59*(6), 839–849.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361–377.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. doi:10.1037/0033-295X.105.2.251

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America*, *104*(4), 2438–2450. doi:10.1121/1.423751

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F., Camus, V., & Bruneau, N. (2013). My voice or yours? An electrophysiological study. *Brain Topography*, *26*(1), 72–82. doi:10.1007/s10548-012-0233-2

Green, P., & Preston, M. (1981). Reinforcement of vocal correlates of auditory hallucinations by auditory feedback: a case study. *The British Journal of Psychiatry*, *139*(3), 204–208. doi:10.1192/bjp.139.3.204

Gur, R. C., & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology*, *37*(2), 147–169. doi:10.1037/0022-3514.37.2.147

Halle, M. (1985). Speculations about the representation of words in memory. *Phonetic Linguistics*, 101–114.

Holzman, P. S., Rousey, C., & Snyder, C. (1966). On listening to one's own voice: Effects on psychophysiological responses and free associations. *Journal of Personality and Social Psychology*, *4*(4), 432–441. doi:10.1037/h0023790

House, A. S., Williams, C. E., Hecker, M. H. L., & Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, *37*(1), 158–166. doi:10.1121/1.1909295

Hughes, S. M., & Nicholson, S. E. (2010). The processing of auditory and visual recognition of self-stimuli. *Consciousness and Cognition*, *19*(4), 1124–1134. doi:10.1016/j.concog.2010.03.001

Jardri, R., Pins, D., Bubrovszky, M., Despretz, P., Pruvo, J.-P., Steinling, M., & Thomas, P. (2007). Self awareness and speech processing: an fMRI study. *NeuroImage*, *35*(4), 1645–1653. doi:10.1016/j.neuroimage.2007.02.002

Johnson, K. (2005). Speaker normalization in speech perception. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 363–389). Cambridge, MA: MIT Press.

Joos, M. (1948). Language Monograph No. 23: Acoustic Phonetics. *Language*, *24*(2), 5–131.

Kaplan, J. T., Aziz-Zadeh, L., Uddin, L. Q., & Iacoboni, M. (2008). The self across the senses: an fMRI study of self-face and self-voice recognition. *Social Cognitive and Affective Neuroscience*, *3*(3), 218–223. doi:10.1093/scan/nsn014

Kawahara, H., & Irino, T. (2005). Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In *Speech Separation by Humans and Machines* (pp. 167–180). Springer US. doi:10.1007/0-387-22794-6_11

Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, *72*(4), 275–309. doi:10.1037/h0021986

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1957). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. doi:10.1037/h0020279

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, *26*(4), 708–715. doi:10.3758/BF03211391

Luce, P. A., & McLennan, C. T. (2005). Spoken word recognition: the challenge of variation. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 591–609). Cambridge, MA: MIT Press.

Magnuson, J. S., Yamada, R. A., & Nusbaum, H. C. (1995). The effects of familiarity with a voice on speech perception. In *Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan* (pp. 391–392).

Massaro, D. W. (1987). *Perceiving talking faces: from speech perception to a behavioral principle*. (S. E. Palmer, Ed.) (p. 507). Cambridge, MA: MIT Press.

Massaro, D. W. (2002). From multisensory information to talking heads. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 153–176). Cambridge, MA: Bradford Books.

Massaro, D. W. (2004). From multisensory information to talking heads. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 153–176). Cambridge, MA: MIT Press.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology:*

Human Perception and Performance, *9*(5), 753–771. doi:10.1037/0096-1523.9.5.753

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. doi:10.1038/264746a0

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*(1), 365–378. doi:10.1121/1.397688

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., & Kojima, S. (2001). Neural substrates for recognition of familiar voices: a PET study. *Neuropsychologia*, *39*(10), 1047–1054.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. doi:10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. doi:10.1111/j.1467-9280.1994.tb00612.x

Olivos, G. (1967). Response delay, psychophysiologic activation, and recognition of one's own voice. *Psychosomatic Medicine*, *29*(5), 433–440.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184. doi:10.1121/1.1906875

Rosa, C., Lassonde, M., Pinard, C., Keenan, J. P., & Belin, P. (2008). Investigations of hemispheric specialization of self-voice recognition. *Brain and Cognition*, *68*(2), 204–214. doi:10.1016/j.bandc.2008.04.007

Rousey, C., & Holzman, P. S. (1967). Recognition of one's own voice. *Journal of Personality and Social Psychology*, *6*(4, Pt.1), 464–466. doi:10.1037/h0024837

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 915–930. doi:10.1037/0278-7393.18.5.915

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), 1447–1469. doi:10.1037//0096-1523.28.6.1447

Shuster, L. I., & Durrant, J. D. (2003). Toward a better understanding of the perception of self-produced speech. *Journal of Communication Disorders*, *36*(1), 1–11. doi:10.1016/S0021-9924(02)00132-6

Smith, R. (2007). The effect of talker familiarity on word segmentation in noise. In *Proceedings of the 16th Congress of the International Phonetics Association* (pp. 1917–1920). Saarbrucken. doi:10.3758/BF03206487

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. doi:10.3758/BF03207704

Strömbergsson, S. (2009). Development of self-voice recognition in children. In *FONETIK* (pp. 136–140).

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters: I. Recognition of backward voices. *Journal of Phonetics*, *13*(1), 19–38.

Chapter 4: Identifying Familiar Speech

4.1 Introduction

When you hear a voice, you identify who is speaking, and you also recognize the words being spoken. Until recently, models of voice perception have assumed these two judgments were based on separate characteristics of a voice, some conveying "identity" and others "speech." To understand speech, a listener extracted the linguistic signals from a voice while disregarding its identifying characteristics. However, it turns out that a voice is not so easily unraveled. Vocal characteristics can contribute to speech, identity, or both (Krieman, Van Lancker-Sidtis, & Gerratt, 2005), and a familiar identity can convey a processing advantage to speech perception (Nygaard, Sommers, & Pisoni, 1994). Consequently, current voice-perception models struggle to determine how speech recognition may be dependent on vocal-identity characteristics. Yet the evidence supporting these models does not indicate whether it is necessary to explicitly identify a talker's voice to gain an advantage from its uniquely-identifying characteristics. Experiments have required listeners to identify talkers, or recognize words, but not perform both tasks simultaneously. We therefore do not know whether a listener must actively identify and recall a talker's voice to generate a familiar-talker advantage, or whether the unique characteristics of an individual voice are automatically encoded and recalled as representations of familiar spoken language, independently of identifying a talker. The present investigation, therefore, required listeners to transcribe speech from

familiar voices as well as identify those voices, to determine whether a speech-processing advantage would be contingent on successfully identifying a talker's unique vocal identity.

4.1.1 Speech and identity as separate processes

Judgments of speech recognition are performed separately from talker identification. These two processes have been identified as neurologically separate and parallel (Knösche, Lattner, Maess, Schauer, & Friederici, 2002; Wood, 1974). Talker identity is processed primarily in the right hemisphere of the brain, and speech in the left (Belin, Fecteau, & Bédard, 2004), although there is overlapping activation between the two (von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003). The speech-processing advantage observed for a familiar voice is particular to the left ear, and thus the right hemisphere (González & McLennan, 2007). Brain-damaged listeners with lesions in their right hemispheres show impaired talker recognition without a loss of speech comprehension (Peretz et al., 1994; van Lancker & Kreiman, 1987), whereas left-hemisphere damage impairs speech comprehension while leaving talker recognition intact (van Lancker & Canter, 1982). Because identity and speech processes appear to proceed separately, but vocal characteristics may be shared between them, an effective model of voice perception must explain how talker-specific characteristics may be used by either process. Current models of speech perception, therefore, focus on how vocal characteristics normally associated with talker identity may be utilized by speech processing.

4.1.2 How talker identity interacts with speech

Abstract models of speech perception maintain that talker-specific characteristics are used to "normalize" speech. That is, in an abstract model, listeners adapt to a talker's identifying characteristics as an internal frame of reference, from which novel speech is more fluently interpreted (Kraljic & Samuel, 2006; Maye, Aslin, & Tanenhaus, 2008; Schweinberger et al., 2008; Schweinberger, Walther, Zäske, & Kovács, 2011; Zäske, Schweinberger, & Kawahara, 2010). The nature of abstract speech forms is a matter of debate (Repp, 1981); they are proposed to be either physical gestures (cf. Fowler, 1986; Liberman & Mattingly, 1985) or a matrix of acoustic features (cf. Ladefoged, 1980). Nonetheless, proponents of abstract models accept that listeners normalize speech into speaker-independent symbolic codes (cf. Bladon, Henton, & Pickering, 1984; Cutler, Eisner, McQueen, & Norris, 2010; Fowler, 2006; Galantucci, Fowler, & Turvey, 2006). In an abstract model, therefore, listeners retain a voice's identifying properties in memory, separately from speech (Green, Tomiak, & Kuhl, 1997), and, upon encountering a previously-heard voice, recall that voice's identifying properties as a reference (cf. Creel & Bregman, 2011). Adapting to talker-specific vocal qualities can therefore facilitate identity judgments (Burton & Bonner, 2004) as well as normalize speech (Dahan, Drucker, & Scarborough, 2008; Eisner & McQueen, 2005; Jesse, McQueen, & Page, 2007; Kraljic & Samuel, 2005). Listeners' adaptation to different vocal identities is observed when processing speech from multiple talkers; the effort required to shift between adaptations slows speech processing and recall (Goldinger, Pisoni, & Logan, 1991; Mullennix, Pisoni, & Martin, 1989; Ryalls & Pisoni, 1997), and it is more effortful

to process speech from multiple talkers than from a single talker (Wong, Nusbaum, & Small, 2004). This evidence may be interpreted as demonstrating reference-frame adaptation because adaptation is not mandatory; a multiple-talker effect does not occur when listeners believe they are hearing a single voice (Johnson, 1990; Magnuson & Nusbaum, 2007) or when speech sounds are easy to recognize, thus lessening the need to adapt to a new talker (McLennan & Luce, 2005). Therefore, if a listener's adaptation to different talker identities is dependent on task and stimulus demands, this would suggest that speech recognition does not require listeners to constantly encode and apply unique reference frames to adapt to every talker they hear. This is concordant with an abstract model, in which talker identity is encoded separately from speech and may be recalled to facilitate normalization of previously-heard talkers. Abstract models of speech perception are therefore a modified version of the traditional view of normalization.

Exemplar models of speech perception propose that we learn and retain talker identity as an integral part of speech tokens. That is, when we are presented with a sound that we know to be speech, we encode the entire acoustic pattern of that *episode* into memory, as a representation of the speech unit (Goldinger, 1998). Listeners, therefore, would encode talker identity directly into speech forms (cf. Pisoni, 1993; Port, 2007; Sheffert & Fowler, 1995; von Kriegstein, Smith, Patterson, Kiebel, & Griffiths, 2010). We thereby accumulate inventories of talker-specific long-term memory traces, wherein each speech unit encodes all the contextual and perceptual details of its original presentation (Hintzman, 1986). Listeners process speech by comparing novel inputs to stored exemplars and selecting a "best match." When speech is being produced by a

known talker, speech processing becomes more efficient because listeners will have a more direct fit for the new input. Therefore, an exemplar model may be supported by evidence showing how identity and speech processes affect each other. For example, the same evidence previously mentioned, purporting to show *detrimental* effects of *multiple* talkers, may instead be interpreted as a *beneficial* effect of a *single* talker (Kaiser, Kirk, Lachs, & Pisoni, 2003), supporting the assertion that listeners may gain a perceptual advantage from encoding talker identity together with speech. Indeed, when a talker's identity and speech are learned together, speech subsequently spoken by that talker is easier to process and remember (Geiselman & Crawley, 1983; Goldinger, 1996). Conversely, learning speech characteristics can make it easier to identify a talker. Certain speech features, especially of nasal vowels, are talker-specific (Amino & Arai, 2007; Amino & Arai, 2009; Marrero et al., 2008); talker-identification accuracy is influenced by language context (Amino & Arai, 2009b) and becomes more accurate when more speech content is present (Bricker & Pruzansky, 1966; Goldstein, Knight, Bailis, & Conover, 1981). Furthermore, when the acoustic characteristics produced by a vocal tract are removed by reducing a voice to sinusoidal replicas, listeners can still identify talkers from the resultant "sine-wave speech" (Fellowes, Remez, & Rubin, 1997; Remez, Fellowes, & Rubin, 1997). These reciprocities between identity and speech can be interpreted as evidence for the integral nature of vocal transmission; that is, in an exemplar model, listeners may learn identity and language as complementary properties of an integrated vocal percept.

4.1.3 A familiar-talker advantage without a familiar identity

Learning to identify a talker has been deemed a "necessary" criterion for receiving a speech-processing advantage from familiar speech. Nygaard, Sommers, & Pisoni (1994) trained listeners over nine days to identify ten different talkers by common English names. At the conclusion of the training period, not all listeners had successfully learned to identify the talkers. When listeners were asked to recognize words spoken by those talkers, "good learners," who had successfully learned to identify the talkers, recognized words more accurately than "poor learners," who had not learned. The authors therefore concluded that identifying a talker was prerequisite to achieving a speech-processing advantage, because listeners who were less capable of identifying talkers did not show speech-processing advantages. However, listeners who are less capable of identifying familiar talkers can still show a speech-processing advantage equivalent to the advantage shown by more-capable listeners (Yonan & Sommers, 2000). Older and younger adults were asked to identify familiar talkers and, subsequently, to recognize the final word of a sentence in noise. Older adults were not as able as young adults to correctly identify familiar talkers, but both age groups demonstrated an equivalent word-recognition advantage from presentation of a familiar voice. Learning to identify a talker by name, therefore, may not be a necessary prerequisite for gaining a familiar-talker advantage. Rather, it may be a correlated effect, such that by the time listeners have learned enough about a voice to reliably identify it, they have also learned enough about that voice's speech patterns to acquire a familiar-talker advantage. The speech-processing advantage gained from learning to identify talkers by name, therefore,

might not be achieved from encoding those talkers' identities for later normalization, but from episodic encoding of their idiosyncratic speech. In this view, "poor learners" do not fail to achieve an advantage because they have failed to learn a talker's identity. Rather, "poor learners" are simply not as able to learn information about a voice.

Indeed, one's ability to identify talkers depends largely on one's ability to encode their speech. Identifying talkers is difficult when the talkers speak an unknown language or meaningless speech sounds (Goggin, Thompson, Strube, & Simental, 1991; Thompson, 1987), and this difficulty can be observed with listeners as young as 7 months of age (Johnson, Westrek, Nazzi, & Cutler, 2011). This difficulty can be ascribed to language content, because identification ability does not improve with repeated exposure to unknown content (Perrachione & Wong, 2007) but instead improves when a listener learns the new language (Schiller & Köster, 1996). Moreover, listeners who are less able to process linguistic content, such as those with dyslexia, find it difficult to identify voices speaking their native language (Perrachione, Del Tufo, & Gabrieli, 2011). The ability to identify a voice therefore appears to be dependent on the ability to encode its speech.

However, the ability to identify a voice cannot be wholly dependent on speech encoding, because listeners can nonetheless learn to identify talkers speaking incomprehensible speech, albeit less accurately than talkers speaking a known language (Schiller, Köster, & Duckworth, 1997). Vocal identities can, in fact, be encoded independently of language. Winters, Levi, & Pisoni (2008) trained monolingual-English listeners to identify bilingual German talkers; each listener was trained either in German

or in English. Listeners were then tested for their ability to identify the same talkers speaking both German and English. English-trained listeners identified English-speaking talkers more accurately than German-speaking talkers. German-trained listeners identified talkers with equivalent accuracy regardless of the language being spoken. Importantly, the English-trained listeners' level of identification accuracy for German-speaking talkers was equivalent to that of the German-trained listeners. This equivalence suggests that all listeners had encoded non-linguistic vocal characteristics, and had been able to use those characteristics to perform identity judgments in the absence of meaningful linguistic content.

Learning a talker's identity could be either a cause or an effect of gaining a familiar-talker speech advantage. That is, a familiar-talker advantage for speech could be obtained through learning to identify a talker, as has been suggested (Nygaard, Sommers, & Pisoni, 1994); alternatively, we could learn to identify a talker through learning to recognize their speech patterns. Speech processing appears to be the more salient judgment when perceiving a voice; in a word-shadowing task, when listeners were paying close attention to speech, 40% of listeners did not notice when the voice they were shadowing switched to a different talker (Vitevich, 2003). When listeners take time to identify voices, they become more aware of different vocal identities and identify them more accurately (Olivos, 1967; Vitevich & Donoso, 2011), suggesting that determinations of talker identity may occur after speech judgments have already been performed. It may be that listeners prioritize each judgment based on the task being performed (Creel & Tumlin, 2011) and, unless talker identification is the primary goal,

speech judgment is the more imperative. This view is supported by the asymmetrical effect of varying talker identity or varying linguistic content. When talker identity is varied between trials, word recognition is detrimentally affected, but when words are varied between trials, talker identification is not as affected (Mullennix & Pisoni, 1990). These observations make it seem less likely that correct identification of a talker is necessary to activate a speech-processing advantage, and introduce the possibility that attention to speech could interfere with identifying a familiar talker. It may be that listeners gain a speech-processing advantage from learning talkers' idiosyncratic speech patterns, and become able to identify talkers as a consequence of learning to identify their speech.

The current debate about talker identity in speech processing centers upon whether talker identity is stored as an inextricable element of speech, or is instead stored separately and recalled to facilitate normalization of abstract speech forms. However, these two models may not be mutually exclusive. It is possible that listeners encode episodic tokens from a voice in addition to separate information about that voice, and implement talker-specific knowledge based on its value to the task being performed. This would mean that, when presented with a speech token spoken by a familiar voice, a listener would not have to explicitly identify the voice to benefit from its familiarity, because the new speech token would match a specific acoustic memory. The question examined by the present study was whether correctly identifying a talker may be significantly related to a speech-processing advantage.

4.1.4 The present study

The present experiment was designed to test whether a familiar-talker word-recognition advantage was contingent on correctly identifying the talker. Previous studies examining the intersection of speech and identity have required participants either to transcribe language, or to identify talkers, but have not required both tasks to be performed simultaneously. Participants in the present investigation transcribed words spoken by familiar and unfamiliar talkers as well as identifying the talker of each word. Words were drawn from phonetically-balanced lists (Egan, 1948). By comparing word-recognition accuracy for when listeners succeeded or failed at identifying familiar talkers, the present results should show to what extent a familiar-talker word-recognition advantage may depend on correct identification.

Participants identified voices as *familiar* or *unfamiliar* from individual speech tokens. Familiarization was induced by passive exposure to two-minute prose passages: "Arthur the Rat" (Abercrombie, 1964) and "The Rainbow Passage" (Fairbanks, 1960). Each passage was approximately two minutes long; passages of this length are adequate to render a talker's speech familiar in both visual and auditory domains (Lander & Davies, 2008; Legge, Grossman, & Pieper, 1984; von Kriegstein et al., 2008). Listeners were not required to recognize talkers by name, because associating a voice with biographical details is an additional, more difficult step than identifying that voice as familiar (Hanley & Damjanovic, 2009). The present experiment featured one familiar voice and three unfamiliar same-sex talkers as foils. Opposite-sex talkers were not used, because they could not be mistaken for a familiar voice; sex classification of an unknown

voice is performed at ceiling levels, even at 4 years of age (Lass et al., 1976; Mullennix, Johnson, Topcu-Durgun, & Farnsworth, 1995; Whiteside, 1998). Familiarity judgments were expected to provide a reliable measure of talker identification.

Participants were trained with either *forward* or *reversed* voices. Reversed voices disrupt talkers' speech, but preserve many vocal qualities, so a talker's reversed voice can be learned and identified, although less accurately (Bartholomeus, 1973; Sheffert, Pisoni, Fellowes, & Remez, 2002; van Lancker, Krieman, & Emmorey, 1985). Listeners were expected to respond to reversed training as they would an unknown language, and learn to identify the talker by voice quality rather than speech-dependent qualities. Listeners who trained with reversed voices were therefore expected to be able to identify normally-speaking familiar talkers in the same way that listeners who learned to identify a bilingual talker speaking an unfamiliar language were able to identify the same talker speaking a familiar language (Winters, Levi, & Pisoni, 2008). Moreover, and more importantly, an abstract model of speech processing proposes that a familiar-talker advantage may be conveyed by encoding a talker's vocal qualities and recalling those qualities to normalize speech. Listeners trained with reversed speech should learn a talker's identifying vocal qualities; therefore, if vocal identity is stored separately from speech, as is indicated by an abstract model, then reverse-trained listeners could also exhibit a familiar-talker advantage.

Every participant was presented with both *normal* and *reversed* words at test, but not all participants transcribed words. Although temporally-reversed words cannot be accurately transcribed, because they do not present valid speech, attempting to transcribe

reversed words should be more effortful than transcribing normally-presented words. The additional effort should slow down listeners' judgments, and slower judgments can enhance listeners' ability to identify talkers by forcing them to pay closer attention (Vitevich & Donoso, 2011). Presenting normal and reversed speech may therefore help disambiguate between an effect of speech content and an effect of paying more attention. If listeners gain an advantage from paying more attention, then listeners who transcribe reversed words should identify talkers more accurately than listeners who do not transcribe words. Alternatively, transcribing words could interfere with talker identification generally. Judgments of speech and identity are separate, and proceed in parallel (Knösche, Lattner, Maess, Schauer, & Friederici, 2002); however, listeners may prioritize these judgments (Creel & Tumlin, 2011). The less-imperative judgment may receive less attention and become less accurate as a consequence. If participants prioritized judgments in this manner, then those performing a word-transcription task would be less accurate at talker identification than participants who did not transcribe, whether the words were presented normally or reversed. A difference in identification performance between transcribers and non-transcribers would support prioritization as a feature of speech processing.

Noise was added to test items as a within-subjects condition, so that words were presented both in noise and in the clear. Noise is commonly used in speech-perception experiments to increase task difficulty and circumvent potential ceiling effects. That is, effects may be seen only more-challenging listening conditions, giving rise to interactions between noise and other factors (Nygaard & Pisoni, 1998).

In sum, for the present experiment, listeners were presented with words spoken by four different same-sex talkers, and one of these four talkers was designated as "familiar." The procedure used two between-subjects factors and two within-subjects factors. Between subjects, participants were either *trained* to be familiar with the designated talker or remained *untrained*, and were either *transcribers* of the words presented or *non-transcribers*. Within subjects, stimuli were presented in *noise* and in the *clear*, and were presented *normally* or temporally *reversed*. Testing was conducted to measure listeners' accuracy both for identifying the familiar talker and for recognizing the words being spoken, and thus explore whether a familiar-talker advantage for word recognition was dependent on being able to correctly identify a talker.

4.2 Experiment

4.2.1 Method

Participants

Seventy participants were recruited (56 female, 14 male, age 18–40 years). All participants spoke English as their first language or had learned English before age 4. Participants were McMaster University students who gave informed consent to the procedures and received course credit for their participation. Each participant was randomly assigned to receive a different type of talker-familiarity training: *forward training*, *backward training*, or *no training*. Each of the trained groups was divided evenly into *transcription* and *no transcription* groups; untrained participants, who would not be able to identify a familiar talker, were all assigned to the *transcription* task.

Members of these groups were pseudorandomly assigned to listen to either male or female talkers, with each sex being heard by an equal number of participants. The number of participants recruited meant that every permutation of between-subjects conditions (e.g., *forward training*, *transcription*, *female*) comprised seven participants. All procedures complied with the tri-council ethics procedures in Canada as approved by the McMaster Research Ethics Board.

Stimuli

Speech stimuli were provided by eight talkers, four male and four female. These talkers were selected from a previous experiment (Chapter 3 here) in which their speech had demonstrated high intelligibility (75% or greater intelligibility in noise). The present experiment used 10 phonetically-balanced sets of 50 words from each talker, as well as two prose passages lasting approximately two minutes each: "Arthur the Rat" and "The Rainbow Passage." Each stimulus was noise-reduced, normalized to a perceived loudness of –17.5 dB, and presented in MP3 format (stereo, 128 kbps, 44100 samples/s). 502 additional stimuli from each talker were created by temporally reversing each normal stimulus, resulting in 1,004 unique stimuli per talker.

A 180-minute white-noise sample, with a constant perceived loudness of –20 dB, was created using Adobe Audition 3 and stored in MP3 format (stereo, 128 kbps, 44100 samples/s).

Apparatus and procedure

Before testing, participants listened to the white noise and were informed that it would be present during half the experiment, at its current loudness, and that its volume

could not be changed. Each participant was solicited to decline participation if they felt that the noise would become intolerable. When a participant indicated their willingness to continue, they were informed that during the experiment, they could remove the headphones and take a break; furthermore, they could change their mind at any time and cease testing with no negative consequence. Although participants were observed to take breaks during the procedure, no participant declined to participate and no participant failed to complete a session.

The testing session was presented via a custom interface, programmed in Realbasic 2008, on a 15" iMac G3 desktop computer running OSX. Audio was played through Maxell HP–200 stereo headphones with the computer's volume fixed at 50% of maximum. All display elements other than the experimental interface were blanked from view.

Prior to testing, *transcription* participants were cautioned that they would not be able to recognize every word. These participants were instructed that if they could not understand a word, they should always type a guess rather than a non-answer (e.g., "didn't hear" or "I don't know"), and informed that typographical errors would be considered incorrect answers.

Each trial presented a single word-token stimulus. In each trial, an alerting asterisk appeared on-screen for 500ms, after which one auditory stimulus was presented. *Transcription* participants transcribed the word, and *no transcription* participants did not. Trained participants then identified the talker by clicking one of two buttons, labeled either "him" and "not him" or "her" and "not her," depending the talkers' sex. Untrained

listeners were presented with buttons marked "A" and "B" and were informed that their choice was irrelevant. After identifying a talker, participants clicked a "next word" button to begin the next trial. A progress bar indicated the proportion of trials completed.

Stimuli were presented in four blocks. Blocks were organized by two factors: *noise* and *stimulus direction*. Stimuli could be presented in the clear or in noise, and could be presented normally or temporally reversed. The four blocks were, therefore, *noise–normal*, *noise–reversed*, *clear–normal*, and *clear–reversed*. Blocks were pseudorandomized so that every possible sequence was presented with equivalent frequency across participants. Each block presented 100 words spoken by four same-sex talkers. Within a block, stimulus selection was random, but constrained so that the same word was not presented more than once. A complete session presented 400 words total.

Each *trained* participant was familiarized with a talker, either male or female, and heard this talker and three unfamiliar sex-matched talkers at test. That is, participants were presented at test with either four male voices or four female voices. Training was accomplished by presenting one prose passage before each of the four blocks. The initial passage was randomly selected and then alternated with the other passage. Participants receiving *forward training* heard normally-presented passages. Participants receiving *backward training* heard temporally-reversed passages. *Trained* participants were instructed to pay attention to each passage, but were informed that they would not be required to remember the passage content. Trained participants thus heard each of the two passages twice during the session, for a total of four exposures.

Untrained participants were each assigned to a particular "familiar" talker, but were not presented with the training passages. Each *untrained* participant therefore was presented with one "familiar" talker and three "unfamiliar" takers, even though all four talkers were unfamiliar.

4.2.2 Results

Results were measured from 70 participants: 28 who had received *forward training*, 28 who had received *backward training*, and 14 *untrained*. Trained groups comprised twice as many participants as untrained because within trained groups, 14 participants performed *transcription* of each stimulus, and 14 performed *no transcription*. Untrained listeners all performed *transcription*. Word-recognition accuracy was measured as the proportion of *normal* stimuli correctly transcribed. *Reversed* stimuli, being incomprehensible nonsense, were excluded from word-accuracy scores. Homonyms were considered correct if no alternative pronunciation were possible (e.g., *rap* vs. *wrap*); however, if an apparent homonym could be disambiguated within a standard Ontario dialect (e.g., *hock* vs. *hawk*), the answer was considered incorrect.

Because processing speech and talker identity are similar processes, both types of judgment could arguably be described as either "recognition" or "identification" interchangeably. To maintain consistency, the present discussion referred to listeners' speech judgments exclusively as word *recognition*, and judgments of talker identity as talker *identification*.

Talker-identification accuracy was measured with the signal-detection formula $d' = Z_{\text{hits}} - Z_{\text{false alarms}}$ (Stanislaw & Todorov, 1999, p. 142). Because Z-scores cannot be

calculated from raw means of 0 or 1, and participants had completed 400 trials, scores of 0 and 1 were adjusted by $\pm\frac{1}{800}$. Untrained listeners, who did not identify talkers, were excluded from analyses of talker identification.

To determine whether listeners had learned to identify the talkers, trained participants' d' scores were compared to a chance level of zero. One-sample t-tests were performed on talker-identification data from *forward training* and *backward training* groups, averaged across noise conditions. Participants who received *forward training* scored greater than chance, $M = 0.87$, $SEM = 0.13$, $t(27) = 6.84$, $p < .001$. Participants who received *backward training* also scored greater than chance, $M = 0.45$, $SEM = 0.08$, $t(27) = 5.40$, $p < .001$. These tests indicated that all trained participants had successfully learned to identify the familiar talkers.

To test whether a familiar-talker advantage for word recognition had resulted from learning to identify the talkers, a mixed-design ANOVA (3 × 2 × 2) was performed on word-recognition data for *transcription* groups (N = 42), using between-subjects factor *training* (forward, backward, untrained), and within-subjects factors *noise level* (clear, noisy), and *talker familiarity* (familiar, unfamiliar). (It will be remembered that untrained listeners, although not familiarized with any talkers, were assigned to the same "familiar" talkers as were the trained groups.) Word-recognition scores were measured as the proportion of words accurately recognized. If training had induced a familiar-talker advantage, this would be observed in an interaction between *training* and *talker familiarity*, because *trained* participants would recognize words more accurately from familiar talkers than unfamiliar talkers, whereas *untrained* participants would show no

advantage for familiar talkers. However, this interaction was not significant, $F(1, 39) = 0.59$, ns. No main effects were observed from either *talker familiarity*, $F(1, 39) = 2.54$, ns, or *training*, $F(2, 39) = 0.01$, ns, although a main effect of *noise level*, $F(1, 39) = 852.11$, $p < .001$, indicated that word recognition was more difficult in noise, $M = .46$, $SEM = .01$, than in the clear, $M = .83$, $SEM = .01$. One interaction was observed, between *noise level* and *talker familiarity*, $F(1, 39) = 45.72$, $p < .001$; this interaction was investigated by performing paired-samples t-tests, comparing *familiar* and *unfamiliar* word-recognition scores in the clear and in noise (Figure 4.1). These tests showed that, in the clear, the intelligibility of *familiar* talkers, $M = .80$, $SEM = .02$, was less than that of *unfamiliar* talkers, $M = .84$, $SEM = .01$, $t(41) = 2.81$, $p = .008$; in noise, the intelligibility of *familiar* talkers, $M = .52$, $SEM = .01$, was greater than that of *unfamiliar* talkers, $M = .44$, $SEM = .01$, $t(41) = -4.84$, $p < .001$. This result cannot be interpreted as a familiar-talker advantage for word recognition in noise, however, because it was not affected by *training*. This result, therefore, may have represented the comparative intelligibility of the individual talkers.

The previous result was surprising, given our *a priori* assumption that familiarization training would successfully convey a word-recognition advantage. To address this assumption, each participant group was tested separately, despite no interactions having been observed between *training* and any other factor. For each group (forward, reverse, untrained), a within-subjects ANOVA was performed (2 × 2) using factors *noise* (clear, noise) and *talker familiarity* (familiar, unfamiliar), specifically examining effects of *talker familiarity*. All three groups showed the same effects. *Talker*

familiarity did not exhibit a main effect: *forward-trained*, $F(1, 13) = 2.00$, *ns*; *reverse-trained*, $F(1, 13) = 1.89$, *ns*; *untrained*, $F(1, 13) = 0.00$, *ns*. A two-way interaction was observed between *talker familiarity* and *noise*: *forward-trained*, $F(1, 13) = 29.50$, $p < .001$; *reverse-trained*, $F(1, 13) = 9.32$, $p = .009$; *untrained*, $F(1, 13) = 12.33$, $p = .004$. Each of these interactions indicated a result similar to that previously described in aggregate: in the clear, familiar talkers were less intelligible than unfamiliar talkers, and in noise, familiar talkers were more intelligible than unfamiliar talkers. All differences were significant except for those of *reverse-trained* listeners in the clear (see Table 4.1). In sum, our *a priori* assumption was erroneous, as the training used in this procedure did not convey a familiar-talker advantage to word recognition.

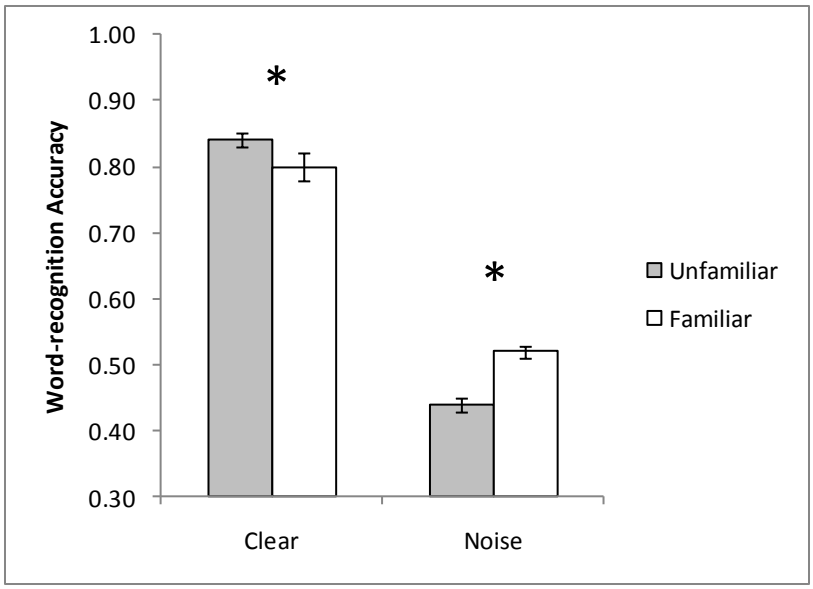

Figure 4.1. Intelligibility of familiar talkers versus unfamiliar talkers in noise and in the clear. Error bars are standard error of the mean corrected for within-subjects design.

Table 4.1. Intelligibility of familiar talkers versus unfamiliar talkers in noise and in the clear, shown by training group.

Training	Noise	Talker	*M*	*SEM*	*t*	df	*p*
Forward	Clear	Familiar	.79	.02	2.16	13	.05
		Unfamiliar	.83	.02			
	Noise	Familiar	.53	.03	–3.58	13	.003
		Unfamiliar	.44	.02			
Reverse	Clear	Familiar	.81	.03	0.53	13	*ns*
		Unfamiliar	.82	.02			
	Noise	Familiar	.52	.01	–2.49	13	.027
		Unfamiliar	.45	.02			
Untrained	Clear	Familiar	.81	.03	2.15	13	.05
		Unfamiliar	.86	.01			
	Noise	Familiar	.50	.02	–2.22	13	.045
		Unfamiliar	.44	.01			

Because training had not produced a familiar-talker advantage, it was not possible to test whether identifying a talker had contributed to a familiar-talker advantage. However, because listeners had learned to identify familiar talkers, it was possible to test whether correctly recognizing words had contributed to talker-identification accuracy. Chi-square tests were performed on contingency tables constructed from individual trials, so as to determine on a trial-by-trial basis whether accurately recognizing a word influenced accurately identifying a talker. Data from untrained listeners, who were unable to identify talkers, were excluded. Contingency tables therefore comprised word recognition and talker identification for each trial, and expected values for each cell were calculated as a proportion of total trials.

For reverse-trained *participants*, no relationship was found between identification and recognition, $X^2(1, N = 2100) = 2.99$, *ns*. For forward-trained participants, however, the relationship was significant, $X^2(1, N = 2100) = 8.78$, $p = .003$. The nature of the relationship was examined by comparing expected values to actual values (Table 4.2), which showed that when forward-trained participants correctly recognized a word, they were more accurate at identifying its talker, and when they failed to correctly recognize a word, they were less accurate at identifying its talker. These tests confirmed that listeners who had learned normal speech were using that speech to help identify a familiar talker.

Table 4.2. Expected and actual values of talker identification for correct and incorrect word recognition. Values have been transformed into proportion of total trials.

		Expected	Actual
Recognized word	Identity correct	.43	.45
	Identity incorrect	.23	.21
Failed to recognize word	Identity correct	.22	.19
	Identity incorrect	.12	.14

It remained to be determined whether either linguistic content or the transcription task had affected talker identification. Identification-accuracy scores were therefore tested for *trained* participants (N = 56) by performing a mixed-design ANOVA (2 × 2 × 2 × 2) using between-subjects factors *training* (forward, backward) and *transcription* (transcription, no transcription) and within-subjects factors *noise level* (clear, noisy) and *stimulus direction* (normal, reversed), looking particularly for effects of *transcription* and *stimulus direction*. If attention to stimulus content had affected talker identification, this would be observed in an effect of *transcription*. If stimulus content itself had affected talker identification, this would be observed in an effect of *stimulus direction*.

Transcription affected talker-identification accuracy. A main effect of *transcription*, $F(1, 52) = 9.21$, $p = .004$, indicated that attention to speech content hindered talker identification; an independent-samples t-test confirmed that participants who transcribed words were less accurate at identifying talkers, $M = 0.45$, $SEM = 0.08$, than participants who did not transcribe words, $M = 0.87$, $SEM = 0.13$, $t(54) = 2.80$, $p = .007$ (Figure 4.2). *Transcription* did not interact with *training*, $F(1, 52) = 0.15$, *ns*, or with *stimulus direction*, $F(1, 52) = 2.38$, *ns*. The failure of *transcription* to interact with *training* or *stimulus direction* indicated that the effect of transcription was due to listeners' attention to stimulus content, and not to the content itself.

Stimulus direction also affected talker-identification accuracy (Figure 4.3). A main effect of *stimulus direction*, $F(1, 52) = 88.16$, $p < .001$, indicated that it was generally more difficult to identify reversed stimuli than normal stimuli. *Stimulus direction* interacted with *noise level*, $F(1, 52) = 7.23$, $p = .01$; paired-samples t-tests,

corrected for multiple comparisons, showed that *normal* stimuli are easier to identify in the clear, $M = 1.27$, $SEM = 0.16$, than in noise, $M = 0.85$, $SEM = 0.13$, $t(55) = 3.48$, $p = .001$, whereas identification accuracy for *reversed* stimuli is equivalent in noise, $M = 0.39$, $SEM = .08$, and in the clear, $M = 0.40$, $SEM = 0.10$, $t(55) = –0.13$, *ns*. The interaction with *training* (Figure 4.4) was tested both within-groups and between-groups. Independent-samples t-tests showed that, between groups, *forward-trained* listeners identified *normal* stimuli more accurately, $M = 1.39$, $SEM = 0.16$, than did *backward-trained* listeners, $M = 0.53$, $SEM = 0.12$, $t(54) = 4.25$, $p < .001$, and that both identified *reversed* stimuli with equivalent accuracy: *forward-trained*, $M = 0.47$, $SEM = 0.11$, *reverse-trained*, $M = 0.34$, $SEM = 0.07$, $t(54) = 0.90$, *ns*. Within-groups paired-samples t-tests indicated that *forward-trained* listeners identified *normal* stimuli more accurately than they did *reversed*, $t(27) = 10.55$, $p < .001$, whereas *backward-trained* listeners were not affected by stimulus direction, showing equivalent accuracy for *normal* stimuli and *reversed* stimuli, $t(27) = 2.17$, ns. The interactions between *stimulus training* and other factors therefore supported the result of the previous chi-square test, showing that *forward-trained* listeners gained an identification advantage when presented with meaningful linguistic content spoken by a familiar talker.

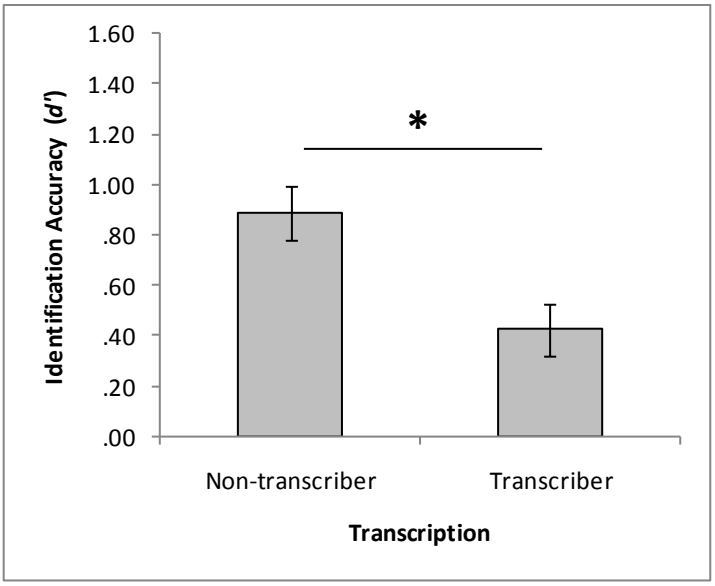

Figure 4.2 Main effect of *transcription* task on talker-identification accuracy. Error bars are standard error of the mean.

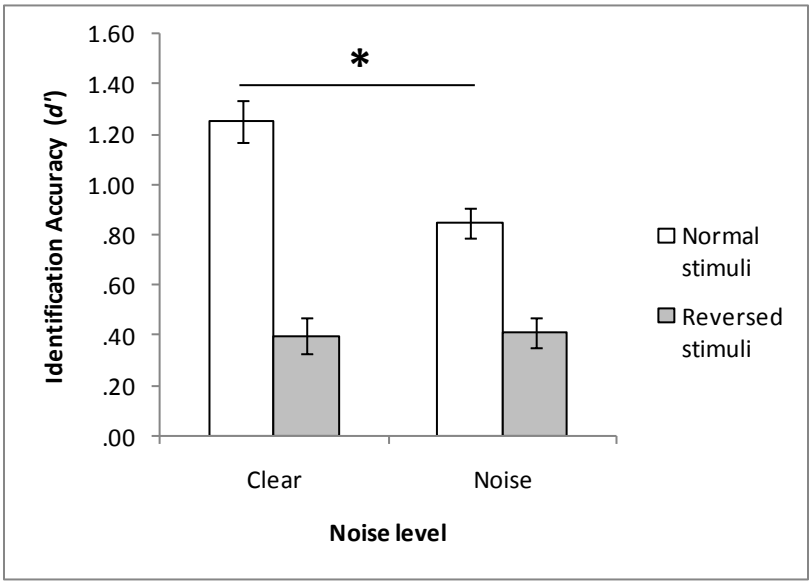

Figure 4.3 Interaction between *stimulus direction* and *noise* affecting talker-identification accuracy. Error bars are standard error of the mean.

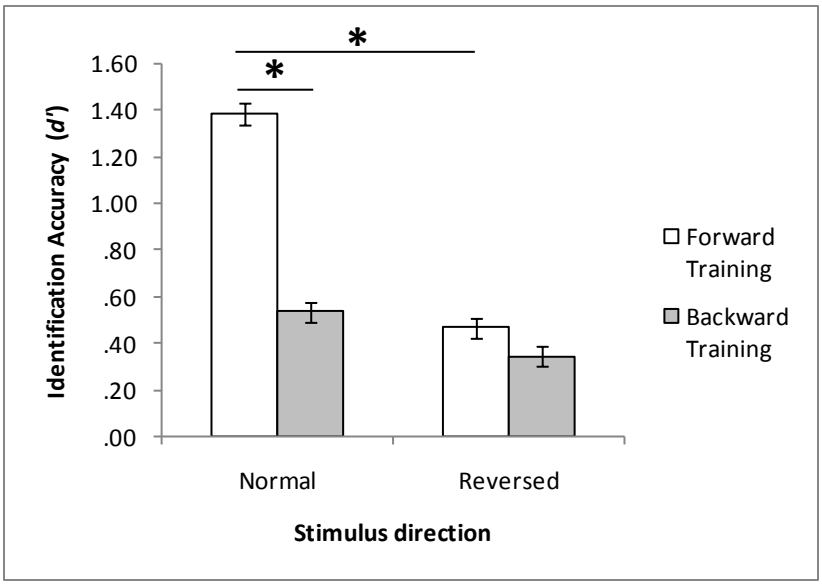

Figure 4.4 Interaction between *stimulus direction* and *training* affecting talker-identification accuracy. Error bars are standard error of the mean corrected for within-subjects design.

4.3 Discussion

Participants listened to prose passages, played either forward or backward, to become familiar with a talker. Listeners then heard words spoken by that familiar talker and others. Words were presented in and out of noise, and were presented normally as well as reversed. Listeners either transcribed or did not transcribe each word before identifying its talker. The results showed that familiar talkers did not convey a speech-processing advantage. Rather, familiar speech conveyed a talker-identification advantage to forward-trained listeners, but not to backward-trained listeners. Participants who performed the transcription task did not show improved talker-identification accuracy; rather, transcription interfered with talker identification, indicating that transcribing listeners prioritized speech judgments over identification. Nonetheless, because listeners who had trained to become familiar with normal speech were able to utilize speech as a clue for performing accurate talker-identity judgments, and listeners who trained with reversed speech could not, these findings support prior evidence showing that talker identity may be encoded as idiosyncratic speech patterns in addition to vocal quality.

4.3.1 The effect of talker identity on speech processing

The primary question of this investigation was whether a familiar-talker advantage for speech processing would be contingent on accurately identifying a familiar talker. This question could not be answered directly, because trained participants did not demonstrate a familiar-talker advantage. Differences between familiar and unfamiliar talkers' intelligibility were observed under certain conditions, but these conditions were not affected by familiarization training; this suggested that the results likely represented

individual differences among the talkers, rather than an effect of familiarity. It must therefore be concluded that the training procedure was insufficient to induce a familiar-talker advantage for speech processing.

4.3.2 The effect of familiar speech on talker identification

The secondary question of this investigation was whether identifying a familiar talker would be affected by speech processing. This was tested by examining effects of word-transcription and reversed-speech stimuli on talker identification. A difference between reversed-speech and normal-speech stimuli would indicate an influence of speech content on talker identification, and an effect of transcription would indicate an influence of attention on talker identification. These effect could have implications for how listeners encode speech sounds when learning a talker's identity and utilize speech sounds when identifying a talker.

Identification accuracy was different for normal-speech versus reversed-speech stimuli, indicating that speech content affected talker identification. Specifically, identification was affected by the presence of speech with which a listener had been familiarized. Listeners had been familiarized either with normal passages which presented valid speech, or reversed passages which did not. Both groups identified reversed stimuli with equivalent accuracy, and reverse-trained listeners also identified normal stimuli at this same level of accuracy. Forward-trained listeners, however, identified normal stimuli more accurately, indicating that forward-trained listeners had been able to utilize familiar speech content to perform more-accurate identification judgments. This result resembles that of monolingual-English listeners who learned to

identify talkers speaking either English or German (Winters, Levi, & Pisoni, 2008). At test, both groups identified German stimuli with equivalent accuracy, and German-trained listeners also identified English stimuli at this same level of accuracy. English-trained listeners, however, identified English stimuli more accurately. The results of both that experiment and this one indicate that listeners who train to recognize a voice will, irrespective of language content, learn certain qualities of a voice that enable them to identify that voice—and that when familiarization training includes recognizable speech, listeners will also encode that speech as an element of the talker's identity, thus becoming able to recognize familiar speech characteristics to facilitate talker identification. Moreover, the present results showed that forward-trained listeners identified familiar talkers more accurately when they correctly recognized words spoken by that talker. The present results therefore provide further evidence that listeners can encode qualities of a talker's identity independently of language, and that listeners can additionally encode a talker's idiosyncratic speech patterns for use in identifying that talker.

Transcribing words made talker identification more difficult. This result would seem to contradict the advantage conveyed by familiar speech, especially in light of prior findings that delayed identification judgments are more accurate (Olivos, 1967; Vitevich & Donoso, 2011). However, the fact that transcription affected all stimuli, and not just familiar talkers, suggests that the difficulty introduced by transcription may be taken as evidence for a model in which identity and speech judgments are prioritized according to task demands (Creel & Tumlin, 2011). Unfortunately, two aspects of the present design rendered this conclusion speculative. One was that the design was not structured as a

dual task; that is, only untrained participants transcribed words without identifying talkers. Because trained participants always identified talkers, the present design could not show the effect of transcription alone. The other aspect was that listeners always transcribed first and identified the talker second, thus making transcription the first-priority task for all participants. No comparison could, therefore, be made versus participants who prioritized identification over transcription. The present results suggest that listeners may have prioritized transcription judgments and performed secondary judgments of talker identity. This possibility should be tested with a procedure designed to do so.

The findings of the present experiment may be related to models of speech processing, despite focusing more on identity processing than speech, because abstract and exemplar models differ in their explication of how talker identity is encoded. An abstract model encodes talker identity separately from speech, whereas an exemplar model encodes talker identity as a component of speech. The present findings support both types of encoding, suggesting that the two models are not necessarily exclusive of each other. A significant relationship between familiar-speech recognition and talker identification was found, which suggests exemplar encoding; however, listeners were able to learn and identify a talker's voice independently of speech, which implies an abstract model. The present results therefore showed that talker-identity encoding was not exclusively bound either to speech forms or to voice quality. Rather, listeners encoded vocal quality as a basis of talker identity and, when a familiar talker's speech was presented in stimuli, utilized speech forms to supplement identification judgments.

The finding that listeners encoded speech both separately from and together with identity suggests, if speculatively, that abstract and exemplar models may not necessarily be in competition, but that speech processing could involve a combination of voice-quality normalization and talker-specific episodic memory.

4.3.3 Present flaws and future directions

The current experimental design could not address its primary question, of whether a familiar-talker advantage was dependent on talker identification, because the familiarization-training procedure was insufficient to induce a familiar-talker advantage. Participants were trained here through passive observation of four two-minute prose passages, which was chosen as an alternative to the nine-day training procedure that first successfully demonstrated a familiar-talker advantage (Nygaard, Sommers, & Pisoni, 1994)—and, in Chapter 2, this two-minute training procedure successfully conveyed a familiar-talker advantage to speech processing. However, the speech being processed in Chapter 2 was a set of only five VCV disyllables. Exposure to two-minute auditory passages may have been adequate to convey an advantage to five predictable consonant sounds, but not to hundreds of monosyllabic words. Exposure to two-minute passages has induced familiar-speech advantages in the visual domain (Lander & Davies, 2008) and the auditory domain (von Kriegstein et al., 2008), but in these cases, all exposures included viewing a talker's dynamic facial movements when speaking. No face was shown to participants in the present procedure. To ensure a familiar-talker advantage, the present procedure should either have replicated Nygaard, Sommers, & Pisoni's nine-day training procedure or, more expediently, trained listeners with exposure to two-minute

audiovisual stimuli. Another possibility, shown to induce an auditory familiar-talker advantage in a single session, would be to require listeners to verbally repeat tokens representative of a talker's speech for approximately one hour (Sanchez, Dias, & Rosenblum, 2013). Successfully inducing a familiar-talker advantage would have made it possible to analyze the effect of talker identity on speech processing.

The present design also could not effectively analyze whether identifying a talker was associated with word-recognition accuracy. The design could therefore be simplified to focus more directly on the question. Three conditions could be eliminated. Participants did not perform at ceiling in the clear, making it unnecessary to present words in noise. Because reversed stimuli did not have an effect other than to make talker identification more difficult, stimuli could all be presented normally. Finally, backward training could be eliminated; although it would be interesting to know whether familiarity with a talker's vocal quality can convey a normalization advantage, that question should be investigated separately. Removing these three factors should increase analytical power and provide a more straightforward treatment of the question.

Four changes would improve the present design. First, training should either be more extensive, or be conducted with audiovisual stimuli, to successfully induce a familiar-talker advantage. Second, a group of trained listeners could perform transcription only, without identifying talkers, making this a dual-task design. Third, the order in which judgments were performed could vary; listeners could either identify a talker and then transcribe a word, or transcribe before identifying. If the effect of transcription on talker identification observed in the present experiment were due to

listeners' prioritization of judgments based on task demands, then listeners should be more accurate for the judgment being performed first. Finally, to produce results that can be generalized, each of the voices featured in the experiment should be familiarized, rather than just one of each sex, so that listeners would serve as controls for each other. Together, these changes should produce results that better address the intended questions.

4.4 Conclusion

The primary findings of the present experiment were that listeners who were familiarized with speech were able to utilize speech to better identify talkers, and that a word-transcription task interfered with talker identification. The theoretical interest of these results was limited, largely because the training procedure failed to induce a familiar-talker advantage for speech processing. Moreover, the complexity and design flaws of the present procedure prevented strong conclusions from being drawn from the analysis of its data. Nonetheless, these findings support a model of voice perception in which a talker's vocal identity is inclusive of its speech production.

4.5 References

Abercrombie, D. (1964). *English phonetic texts*. London: Faber and Faber.

Amino, K., & Arai, T. (2007). Contribution of consonants and vowels to the perception of speaker identity. In *Japan-China Joint Conference of Acoustics* (p. CD–ROM). Sendai, Japan.

Amino, K., & Arai, T. (2009a). Speaker-dependent characteristics of the nasals. *Forensic Science International*, *185*(1-3), 21–28. doi:10.1016/j.forsciint.2008.11.018

Amino, K., & Arai, T. (2009b). Effects of linguistic contents on perceptual speaker identification: Comparison of familiar and unknown speaker identifications. *Acoustical Science and Technology*, *30*(2), 89–99. doi:10.1250/ast.30.89

Bartholomeus, B. (1973). Voice identification by nursery school children. *Canadian Journal of Psychology*, *27*(4), 464–472. doi:10.1037/h0082498

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends In Cognitive Sciences*, *8*(3), 129–135. doi:10.1016/j.tics.2004.01.008

Bladon, R. A. W., Henton, C. G., & Pickering, J. B. (1984). Towards an auditory theory of speaker normalization. *Language & Communication*, *4*(1), 59–69. doi:10.1016/0271-5309(84)90019-3

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America*, *40*(6), 1441–1449. doi:10.1121/1.1910246

Burton, A. M., & Bonner, L. (2004). Familiarity influences judgments of sex: The case of voice recognition. *Perception*, *33*(6), 747–752. doi:10.1068/p3458

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass*, *5*(5), 190–204. doi:10.1111/j.1749-818X.2011.00276.x

Creel, S. C., & Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language*, *65*(3), 264–285. doi:10.1016/j.jml.2011.06.005

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory Phonology 10* (pp. 91–111). Berlin: de Gruyter.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: adjusting the signal or the representations? *Cognition*, *108*(3), 710–718. doi:10.1016/j.cognition.2008.06.003

Egan, J. P. (1948). Articulation testing methods. *The Laryngoscope*, *58*(9), 955–991. doi:10.1288/00005537-194809000-00002

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. doi:10.3758/BF03206487

Fairbanks, G. (1960). *Voice and Articulation Drillbook* (2nd ed., p. 234). New York, NY: Harper & Brothers.

Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics*, *59*(6), 839–849. doi:10.3758/BF03205502

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*(1), 3–28.

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, *68*(2), 161–177. doi:10.3758/BF03193666

Galantucci, B., Fowler, C. a, & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361–377. doi:10.3758/BF03193857

Geiselman, R. E., & Crawley, J. M. (1983). Incidental processing of speaker characteristics: voice as connotative information. *Journal of Verbal Learning and Verbal Behavior*, *22*(1), 15–23. doi:10.1016/S0022-5371(83)80003-6

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. doi:10.3758/BF03199567

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183. doi:10.1037/0278-7393.22.5.1166

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. doi:10.1037/0033-295X.105.2.251

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 152–162. doi:10.1037/0278-7393.17.1.152

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.

González, J., & McLennan, C. T. (2007). Hemispheric differences in indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 410–424. doi:10.1037/0096-1523.33.2.410

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception & Psychophysics*, *59*(5), 675–692. doi:10.3758/BF03206015

Hanley, J. R., & Damjanovic, L. (2009). It is more difficult to retrieve a familiar person's name and occupation from their voice than from their blurred face. *Memory*, *17*(8), 830–839. doi:10.1080/09658210903264175

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428. doi:10.1037//0033-295X.93.4.411

Jesse, A., McQueen, J. M., & Page, M. (2007). The locus of talker-specific effects in spoken word recognition. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1921–1924). Dudweiler: Pirrot.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–1011. doi:10.1111/j.1467-7687.2011.01052.x

Johnson, K. A. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, *88*(2), 642–654. doi:10.1121/1.399767

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *46*(2), 390–404. doi:10.1044/1092-4388(2003/032)

Knösche, T. R., Lattner, S., Maess, B., Schauer, M., & Friederici, A. D. (2002). Early parallel processing of auditory word and voice information. *Neuroimage*, *17*(3), 1493–1503. doi:10.1006/nimg.2002.1262

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178. doi:10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268. doi:10.3758/BF03193841

Kreiman, J., van Lancker, D. R., & Gerratt, B. R. (2005). Perception of voice quality. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 338–362). Malden, MA: Blackwell Publishing.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language*, *56*(3), 485–502.

Lander, K., & Davies, R. (2008). Does face familiarity influence speechreadability? *Quarterly Journal of Experimental Psychology*, *61*(7), 961–967. doi:10.1080/17470210801908476

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. *Journal of the Acoustical Society of America*, *59*(3), 675–678. doi:10.1121/1.380917

Legge, G. E., Grosmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 298–303. doi:10.1037/0278-7393.10.2.298

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409. doi:10.1037/0096-1523.33.2.391

Marrero, V., Battaner, E., Gil, J., Llisterri, J., Machuca, M., Marquina, M., De La Mota, C., & Rios, A. (2008). Identifying speaker-dependent acoustic parameters in Spanish vowels. *Journal of the Acoustical Society of America*, *123*(5), 3877. doi:10.1121/1.2935781

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562. doi:10.1080/03640210802035357

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, *31*(2), 306–321. doi:10.1037/0278-7393.31.2.306

Mullennix, J. W., Johnson, K. A., Topcu-Durgun, M., & Farnsworth, L. M. (1995). The perceptual representation of voice gender. *Journal of the Acoustical Society of America*, *98*(6), 3080–3095. doi:10.1121/1.413832

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390. doi:10.3758/BF03210878

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*(1), 365–378. doi:10.1121/1.397688

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. doi:10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. doi:10.1111/j.1467-9280.1994.tb00612.x

Olivos, G. (1967). Response delay, psychophysiologic activation, and recognition of one's own voice. *Psychosomatic Medicine*, *29*(5), 433–440.

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., & Belleville, S. (1994). Functional dissociations following bilateral lesions of auditory cortex. *Brain*, *117*(6), 1283–1301. doi:10.1093/brain/117.6.1283

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899–1910. doi:10.1016/j.neuropsychologia.2006.11.015

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, *13*(1-2), 109–125. doi:10.1016/0167-6393(93)90063-Q

Port, R. F. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas In Psychology*, *25*(2), 143–170. doi:10.1016/j.newideapsych.2007.02.001

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 651–666. doi:10.1037/0096-1523.23.3.651

Repp, B. H. (1981). On levels of description in speech research. *Journal of the Acoustical Society of America*, *69*(5), 1462–1464. doi:10.1121/1.385779

Ryalls, B. O., & Pisoni, D. B. (1997). The effect of talker variability on word recognition in preschool children. *Developmental Psychology*, *33*(3), 441–452.

Sanchez, K., Dias, J. W., & Rosenblum, L. D. (2013). Experience with a talker can transfer across modalities to facilitate lipreading. *Attention, Perception, & Psychophysics*, *75*(7), 1359–1365. doi:10.3758/s13414-013-0534-x

Schiller, N. O., Köster, O., & Duckworth, M. (1997). The effect of removing linguistic information upon identifying speakers of a foreign language. *Forensic Linguistics*, *4*(1), 1–17.

Schweinberger, S. R., Casper, C., Hauthal, N., Kaufmann, J. M., Kawahara, H., Kloth, N., Robertson, D.M.C., Simpson, A.P., & Zäske, R. (2008). Auditory adaptation in voice perception. *Current Biology*, *18*(9), 684–688. doi:10.1016/j.cub.2008.04.015

Schweinberger, S. R., Walther, C., Zäske, R., & Kovács, G. (2011). Neural correlates of adaptation to voice identity. *British Journal of Psychology*, *102*(4), 748–764. doi:10.1111/j.2044-8295.2011.02048.x

Sheffert, S. M., & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory and Language*, *34*(5), 665–685. doi:10.1006/jmla.1995.1030

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), 1447–1469. doi:10.1037/0096-1523.28.6.1447

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*(2), 121–131. doi:10.1002/acp.2350010205

Van Lancker, D. R., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition*, *1*(2), 185–195. doi:10.1016/0278-2626(82)90016-1

Van Lancker, D. R., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829–834. doi:10.1016/0028-3932(87)90120-5

Van Lancker, D. R., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters: I. Recognition of backward voices. *Journal of Phonetics*, *13*(1), 19–38.

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 333–342. doi:10.1037/0096-1523.29.2.333

Vitevitch, M. S., & Donoso, A. (2011). Processing of indexical information requires time: Evidence from change deafness. *Quarterly Journal of Experimental Psychology*, *64*(8), 1484–1493. doi:10.1080/17470218.2011.578749

Von Kriegstein, K., Dogan, O., Grüter, M., Giraud, A., Kell, C. A., Grüter, T., Kleinschmidt, A., & Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences*, *105*(18), 6747–6752. doi:10.1073/pnas.0710826105

Von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A.-L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*(1), 48–55. doi:10.1016/S0926-6410(03)00079-X

Von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience*, *30*(2), 629–638. doi:10.1523/JNEUROSCI.2742-09.2010

Whiteside, S. P. (1998). Identification of a speaker's sex: a study of vowels. *Perceptual and Motor Skills*, *86*(2), 579–584. doi:10.2466/pms.1998.86.2.579

Winters, S. J., Levi, S. V, & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America*, *123*(6), 4524–4538. doi:10.1121/1.2913046

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience*, *16*(7), 1173–1184. doi:10.1162/0898929041920522

Wood, C. C. (1974). Parallel processing of auditory and phonetic information in speech discrimination. *Perception & Psychophysics*, *15*(3), 501–508. doi:10.3758/BF03199292

Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging*, *15*(1), 88–99. doi:10.1037/0882-7974.15.1.88

Zäske, R., Schweinberger, S. R., & Kawahara, H. (2010). Voice aftereffects of adaptation to speaker identity. *Hearing Research*, *268*(1-2), 38–45. doi:10.1016/j.heares.2010.04.011

Chapter 5: General Discussion

5.1 Thesis Summary

The primary goal of this thesis was to explore how we relate a talker's vocal identity to the speech he or she produces. Until recently, variability among talkers' voices was considered a perceptual problem. That is, speech perception was modeled as a process by which listeners neutralized individuals' vocal characteristics to extract standardized speech signals (cf. Joos, 1948). This model was uprooted by the demonstration of a *familiar-talker advantage* for speech processing (Nygaard, Sommers, & Pisoni, 1994), which showed that individuals' vocal characteristics are not neutralized, but may be retained and recalled to facilitate speech processing. The familiar-talker advantage therefore called into question our understanding of how vocal identity was stored in memory in relation to speech. Two primary hypotheses emerged to explain the relationship between vocal identity and speech perception: the *normalization model* (cf. Norris, 1994), in which vocal identity is stored in memory and used as a frame of reference when processing speech from that talker; and the *exemplar model* (cf. Goldinger, 1998), in which speech sounds are stored as talker-specific exemplars, and each incoming signal is processed by making a "best fit" to the inventory of stored sounds. With these hypotheses in view, the present investigations used familiar voices to explore how talker identity may be encoded and recalled in speech processing.

Chapter 2 used the *McGurk effect* to determine which of a familiar face or voice contributed more strongly to audiovisual speech perception. It is the first experiment to have made such a comparison. The principle of optimal integration predicted that a familiar face or voice would produce greater weighting in its respective channel, and disrupt the McGurk illusion in favor of visual or auditory percepts, respectively; therefore, listeners' visual or auditory responses were expected to measure the relative contributions of familiar faces and voices. However, not all individuals were equally affected by the illusion. Individual differences between listeners drove us to categorize them as either "high" or "low" integrators (i.e., more or less efficient at integrating audiovisual speech). High integrators used familiar faces to integrate more efficiently, and low integrators used familiar faces and voices to reject incongruous speech signals more effectively. Listeners' integration levels were not attributable to their personal sensory biases, but to familiarization training: participants who did not undergo training were biased toward auditory weighting, whereas trained listeners weighted the visual channel more heavily. Nonetheless, regardless of weighting or training, a familiar-talker advantage was activated only by the presentation of a familiar voice, and not a familiar face, indicating that a familiar voice has a stronger effect on audiovisual speech perception in the context of the McGurk illusion.

Chapter 3 tested whether self-voice conveyed a familiar-talker advantage, and is the first experiment to have done so. Participants recorded lists of words and then, in a second session, transcribed words spoken by themselves and others. A self-talker advantage was observed, but was affected by whether stimuli were presented in noise or

in the clear. When stimuli were presented in noise, listeners showed an equivalent word-recognition advantage for all voices of their own sex, including self-voice. When stimuli were presented in the clear, listeners showed an advantage for only self-voice. These results confirm that listeners demonstrate a speech-processing advantage for their own voices, and present the new finding that a speech-processing advantage can be conveyed by unfamiliar voices similar to one's own.

Chapter 4 sought to determine whether a familiar-talker advantage was dependent on explicitly identifying a voice. Unfortunately, the experiment could not answer its original question, because training failed to induce a familiar-talker advantage for word recognition. However, participants did learn to identify familiar talkers, and two effects of speech processing on talker identity were observed. For one, actively transcribing words made talker identification more difficult. For another, all trained listeners became able to identify talkers, but listeners who had trained with normal speech instead of reversed speech received an identification advantage from hearing normal speech versus reversed speech. These results suggest that listeners encode a talker's idiosyncratic speech production as part of a talker's vocal identity, and may recall what they have encoded to facilitate talker-identification judgments.

Overall, the findings of this investigation present new information about the relation between a familiar talker and speech processing. Chapter 2 demonstrated that a familiar voice is more important than a familiar face to audiovisual speech processing. Chapter 3 expanded on the nature of a familiar-talker advantage by showing that familiar-sounding unfamiliar voices can convey a speech-processing advantage. Chapter 4

contributed new evidence to show how listeners encode speech production as interrelated with talker identity.

5.2 Implications and Significance

The present findings have implications for how we encode and recall talker identity in relation to speech. Current models of speech processing suggest either that qualities of talker identity are encoded alongside speech and recalled as a frame of reference to normalize abstract speech forms from that talker, or that talker-identifying characteristics are encoded as speech and stored as a talker-specific inventory of episodic exemplars. The present findings inform these models, firstly, by demonstrating individual differences in listeners' encoding and recall of talker-identifying information; secondly, by indicating that the utility of familiar vocal characteristics in speech processing is not necessarily talker-specific; and thirdly, by suggesting that encoding of talker identity includes a talker's idiosyncratic speech production. All together, these findings suggest that the two primary models of speech perception currently being debated may not be exclusive of each other, but may instead describe different aspects of talker encoding and recall.

5.2.1 Familiar audiovisual speech

A familiar talker's speech may be encoded visually (face) or aurally (voice). We know that listeners can recall what they have encoded to facilitate processing that same talker's speech, either in the same mode (Lander & Davies, 2008; Nygaard, Sommers, & Pisoni, 1994) or in the other mode (Rosenblum, Miller, & Sanchez, 2007; Sanchez, Dias,

& Rosenblum, 2013). These data, however, do not represent audiovisual speech perception. Chapter 2 demonstrated that, in audiovisual speech presentation, listeners' recall of familiar talkers was activated by a familiar voice and not by a familiar face. A familiar-talker advantage for visual speech has been demonstrated (Lander & Davies, 2008; Sanchez, Dias, & Rosenblum, 2013), but only in silent lip-reading tasks. In the present experiment, when both auditory and visual signals were present, no visual-speech advantage obtained. This result suggests that, regardless of the mode in which a talker's speech was made familiar, listeners are biased to activate a familiar-talker advantage in response to an auditory signal.

Listeners in Chapter 2, however, were not necessarily biased toward audition. Optimal integration predicts that a sensory channel with greater reliability will be weighted more heavily; it was therefore predicted that a familiar face or voice would produce greater weighting in its respective channel. This did not occur. Noting that listeners' responses were influenced by presentation of a familiar voice, and not a familiar face, optimal integration might have predicted that listeners would weight the auditory channel more heavily. This also did not occur. Instead, familiarization training caused listeners to weight the visual channel more heavily, even when responding to a familiar voice. This result implies that the mechanism of optimal integration may work differently than has been previously represented. Weightings in optimal integration have, before now, been thought attributable solely to the objective "reliability" of a channel; that is, to the relative physical signal strength or clarity of one channel versus another. The present results show that increasing the perceptual reliability of a sensory channel

did not necessarily increase weighting to that channel. The current findings therefore indicate that, independently of a channel's objective reliability, a listener may subjectively decide that one channel is more important than another and, in so doing, attribute more weight to that channel.

Chapter 2 observed that a familiar-talker advantage did not affect participants' overall performance as either "low integrators" or "high integrators." Differences in integration performances have been observed before, and are hypothesized to exist either from individual differences in integration skill (Grant, 2002) or from individuals' natural biases toward one sensory channel which reduces integration through undue weighting (Schwartz, 2010). The findings of the present experiment favor the former hypothesis of individual differences in skill level. Individuals did demonstrate biases for either visual or auditory channels, but these biases were inculcated by training and did not determine an individual's overall level of integration performance. Different levels of integration skill could explain why a familiar-talker advantage enabled high integrators to integrate more efficiently and low integrators to fixate more closely on a single channel. However, this explanation challenges a central precept of optimal integration. Optimal integration states that increasing the reliability of one channel increases "weighting" in that channel, increasing the likelihood of that channel being perceived instead of an integrated percept. In the present experiment, this did not occur. Rather, increasing the reliability of one channel enhanced listeners' natural tendency to integrate or de-integrate a stimulus. This result indicates that optimal integration may not be driven solely by relative channel reliability, but can be subject to individuals' strategic differences.

5.2.2 Familiar self-speech

Prior to the present investigation, we did not know whether recordings of self-voice would convey a familiar-talker advantage to speech processing, even though self-voice recordings were recognized as a familiar voice (Kaplan, Aziz-Zadeh, Uddin, & Iacoboni, 2008). The data presented in Chapter 3 are the first to confirm a familiar-talker advantage for self speech, but in doing so raise additional questions. For one, we cannot be sure from these data from what source self-voice is encoded. Our own recorded voice does not sound the same to us as our normal speaking voice, because our voice normally reaches our ears by bone conduction as well as through the air (Békésy, 1949). Self-voice could therefore be encoded as the bone-conducted sound as well as the recorded sound, meaning that the familiar-talker advantage conveyed by self-voice recordings might be attributable to encoding one's own idiosyncratic speech production from bone-conducted speech, even though its sound quality is different from a recording (Shuster & Durrant, 2003). On the other hand, a voice's sound quality was salient to the present results. When voices' sound quality was obscured by noise, listeners demonstrated a speech-processing advantage for all voices exhibiting the same sex qualities as their own voice. This raises the question of whether listeners were simply orienting themselves to a sex-typical reference frame (Johnson, 2005) or were instead recalling certain characteristics of self-voice that conveyed an advantage to unfamiliar voices.

Research has not addressed whether a talker's voice may be "partially" recalled to facilitate speech processing, as appeared to occur in Chapter 3. In Chapter 3, when listening conditions were made difficult, listeners gained a speech-processing advantage

not only for their own voices, but for all voices that sounded similar to their own, suggesting that listeners had been able to recall and apply those similar characteristics to facilitate speech processing, even though other characteristics of self-voice did not match. An exemplar model recognizes that there are levels of encoding, such that a listener will remember whichever vocal elements he or she attends to at the time of its first presentation, but does not explicitly describe what happens when an initial encoding is of a superior resolution to subsequent presentations of a voice (cf. Goldinger, 1998). Similarly, discussion of abstract models acknowledges both an effect of individual talkers and an effect of general reference frames (cf. Johnson, 2005), but has not explicitly considered a situation in which listeners adapt to a general reference frame derived from one individual talker. Both exemplar and abstract models are theoretically capable of accommodating this new evidence; the current findings mainly imply that each model should no longer consider talker encoding as a black-box bundle of characteristics that, when recalled in speech processing, is recalled *in toto*. Instead, each model should ask to what extent, and under what conditions, elements of a talker's encoded voice will be recalled for use in speech processing.

5.2.3 The role of talker identification

Chapter 4 provides evidence that, when learning to identify a talker, listeners are not limited to encoding acoustic qualities of a talker's voice, but also encode a talker's manner of speech production. The present investigation was not the first procedure to train listeners using backwards speech (Sheffert, Pisoni, Fellowes, & Remez, 2002), nor is it the first to test listeners with reversed stimuli (Goggin, Thompson, Strube, &

Simental, 1991), but it was the first to test backward-trained listeners using reversed stimuli. The present findings were similar to cross-language effects of talker identification (Winters, Levi, & Pisoni, 2008), such that all listeners identified voices equally well when making judgments based on vocal quality alone, and that listeners who had been trained through learning recognizable speech were able to use that learning to identify talkers more accurately. This meant that forward-trained listeners had indeed become familiar with talkers' speech, although not to the extent of gaining a familiar-talker advantage for speech processing. Chapter 4's results therefore may imply a challenge to the assertion that learning a familiar identity necessarily activates a speech-processing benefit (Nygaard, Sommers, & Pisoni, 1994). Rather, the present results suggest that listeners may gain a familiar-talker benefit from learning a talker's idiosyncratic speech production, and also learn identity as a separate effect. The possibility of listeners gaining a speech-processing from learning speech, rather than identity, would not exclude the converse situation in which a talker's identity, once learned, may be used to activate relevant memory stores; however, it is a possibility not currently being accommodated by models that address the intersection of talker identity and speech perception.

5.3 Limitations and Future Directions

Each of the experiments suffered from complexity. Whether an overabundance of stimuli, a plethora of factors, or unneeded conditions, results were generally less powerful and findings less clear than they would have been if procedures had been scaled

back. Future directions for this work therefore include suggestions for similar but more-focused experiments as well as recommended next steps.

Chapter 2's familiar-McGurk experiment could have presented fewer talkers than the nineteen that were featured. The number of talkers used meant that the familiar talker was made obvious to participants as a talker of interest by that talker's appearance in fifty percent of all trials, which in turn may have generated a learning effect, obscuring the effect of familiar versus unfamiliar talkers within the "low integrator" participant group. The existence of a learning effect would argue against a procedure with only two talkers, because the unfamiliar talker would be present 50% of the time and could therefore be learned from; but, historically, experiments using the McGurk effect have found significant results with as few as four talkers. Overexposure of the familiar talker may also have made it too easy for listeners to match faces and voices; without such matching, perhaps visually-trained listeners would not have responded so strongly to familiar voices. The relative effects of familiar faces and voices might appear differently if talker-identity matching were implemented as a between-subjects factor, such that half the participants would be presented only with matched-identity stimuli and the rest only with mismatched-identity stimuli. With or without this manipulation, though, reducing the quantity of talkers in the present procedure should clarify familiar-talker effects.

A follow-up experiment to Chapter 2 could explore the apparent association between listeners' modal biases and their demonstrated "skill level" in audiovisual integration. To test whether listeners' modal biases were expressed in encoding, and not in response, listeners could be trained as in the present procedure, but the familiar talker

not presented in testing. If listeners were "naturally" biased toward auditory or visual modes then, in the absence of a modal-orienting familiar talker, their responses should reflect those biases, regardless of training mode (Schwartz, 2010). To test the present findings that associate high and low integration "skill" with visual and auditory learning, respectively, a series of experiments could test listeners for their baseline level of audiovisual integration and then apply a diagnostic test of their learning style (e.g., Riding, 1991). These tests could inform the current discussion regarding why listeners demonstrated different levels of audiovisual-integration ability.

The number of conditions implemented in Chapter 3 had a negative effect on statistical power. Eight talkers and four levels of noise made it difficult to disambiguate effects of same-sex voices versus those of self-voice. Experiment 2 removed four talkers and presented only two levels of noise, but added the complication of reversed stimuli. An experiment featuring two same-sex talkers, two levels of noise, and only normal stimuli would produce greater power.

Although Chapter 3 did answer the question of whether self-voice conveys a familiar-talker advantage, it left open the question of why self-voice conveyed an equivalent advantage to same-sex voices. When stimuli were presented in noise, listeners may either have switched to a same-sex orientation or, being oriented to the sex-typical qualities of their own voices, processed nonself speech by reference to those same encoded qualities. Gender coding has been observed to influence speech processing (Church & Schacter, 1994), but the present result also echoes the general advantage conveyed to speech processing by audiovisual self-speech (Aruffo & Shore, 2012).

Chapter 3's result could be tested with a follow-up experiment presenting listeners with a familiar talker, in noise, along with opposite-sex voices sharing overtly-similar characteristics (e.g., fundamental speaking pitch, regional accent, intonation). Measuring whether the same advantage conveyed by the familiar voice extended to these similar-sounding voices could determine whether or not a general advantage could be derived from a single talker.

Chapter 4's familiar-speech procedure could have been performed with fewer factors. The training procedure's failure to inculcate a familiar-talker advantage diminished the design's ability to examine the relationship between talker identity and word recognition, but the multitude of factors weakened power, making it difficult to interpret the results that were found. It was also a procedural error to familiarize all listeners with the same two talkers, instead of balancing all the talkers among listeners. This experiment could be re-run with a more-effective training procedure (e.g., Sanchez, Dias, & Rosenblum, 2013), using only four talkers, while eliminating the factors of *noise*, *stimulus direction*, and *talker sex,* as well as backward training. The design could then be easily modified to a dual-task design by allocating a group of participants to undergo familiarization training but only transcribe words at test.

The next step for a familiar-speech study would depend on whether the revised experiment showed an association between correct talker identification and greater word-recognition accuracy. If no association were found, then it might be concluded that learning a talker's identity is incidental to obtaining a familiar-talker speech advantage. If an association were found, however, the next step would be to ascertain whether the

speech-processing advantage was dependent on correctly identifying a talker. This could be achieved by morphing a familiar voice so that it could no longer be identified as the original talker, but to do so without altering the talker's temporal speech characteristics. For example, the overall pitch of a voice could be raised to an unfamiliar range without altering the timing of that voice's articulatory actions. A familiar-talker advantage would likely be lessened by such a manipulation, because its familiar frequency referents would have been raised, but the voice's familiar articulations could nonetheless provide an advantage versus unfamiliar speech. The appearance or non-appearance of such an advantage would contribute evidence toward the necessity or inessentiality of talker identification in speech processing.

5.4 Conclusion

Speech-perception models are currently challenged to determine how listeners encode and recall talker-specific vocal information. The earliest view, of talker information being irrelevant and discarded, became untenable once it was clear that speech processing benefited from same-voice repetition. However, it is now unclear whether talker identity is encoded separately from speech forms or integrally with them, and what vocal characteristics will be recognized as "identity" or "speech." This thesis contributes to this discussion by studying familiar talkers, whose vocal characteristics listeners have encoded into memory and have available for recall. Chapter 2, an exploration of audiovisual speech processing, showed that listeners may be biased to encode either a talker's face or voice, but then recall what they have encoded when

processing a familiar voice, and not a face. Chapter 3, a study of self-speech perception, demonstrated an association between talker identification and word recognition, suggesting that listeners may recall a partial representation of a familiar talker's voice when hearing an impoverished signal. Chapter 4, an examination of identifying talkers from familiar speech, found that familiar speech was recalled to facilitate talker identification and not speech processing. Together, these findings add new evidence and introduce new perspectives to the discussion of how talker identities are encoded into memory and recalled in speech processing.

5.5 References

Aruffo, C., & Shore, D. I. (2012). Can you McGurk yourself? Self-face and self-voice in audiovisual speech. *Psychonomic Bulletin & Review*, *19*(1), 66–72. doi:10.3758/s13423-011-0176-8

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110–120. doi:10.1007/s00426-008-0185-z

Békésy, G. V. (1949). The structure of the middle ear and the hearing of one's own voice by bone conduction. *Journal of the Acoustical Society of America*, *21*(3), 217–232. doi:10.1121/1.1906501

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. doi:10.1016/j.cognition.2007.04.005

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, *61*(2), 206–219. doi:10.3758/BF03206883

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America*, *40*(6), 1441–1449. doi:10.1121/1.1910246

Carr, P. B., & Trill, D. (1964). Long-term larynx-excitation spectra. *Journal of the Acoustical Society of America*, *36*(11), 2033–2040. doi:10.1121/1.1919319

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 521–533. doi:10.1037/0278-7393.20.3.521

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior*, *4*(4), 373–394. doi:10.1007/BF01040628

Cole, R. A., Coltheart, M., & Allard, F. (1974). Memory of a speaker's voice: reaction time to same- or different-voiced letters. *Quarterly Journal of Experimental Psychology*, *26*(1), 1–7. doi:10.1080/14640747408400381

Craik, F. I. M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, *26*(2), 274–284. doi:10.1080/14640747408400413

Doddington, G. R. (1985). Speaker recognition—Identifying people by their voices. *Proceedings of the IEEE*, *73*(11), 1651–1664. doi:10.1109/PROC.1985.13345

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends In Cognitive Sciences*, *8*(4), 162–169. doi:10.1016/j.tics.2004.02.002

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, *16*(1), 78–80. doi:10.1109/TAU.1968.1161953

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. doi:10.3758/BF03199567

Goh, W. D. (2005). Talker variability and recognition memory: instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 40–53. doi:10.1037/0278-7393.31.1.40

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. doi:10.1037/0033-295X.105.2.251

Halle, M. (1985). Speculations about the representation of words in memory. *Phonetic Linguistics*, 101–114.

Hanley, J. R., & Damjanovic, L. (2009). It is more difficult to retrieve a familiar person's name and occupation from their voice than from their blurred face. *Memory*, *17*(8), 830–839. doi:10.1080/09658210903264175

Harnsberger, J. D., Shrivastav, R., Brown, W. S., Rothman, H., & Hollien, H. (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice*, *22*(1), 58–69. doi:10.1016/j.jvoice.2006.07.004

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*(5), 3099–3111. doi:10.1121/1.411872

Hughes, S. M., & Nicholson, S. E. (2010). The processing of auditory and visual recognition of self-stimuli. *Consciousness and Cognition*, *19*(4), 1124–1134. doi:10.1016/j.concog.2010.03.001

Johnson, K. (2005). Speaker normalization in speech perception. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 363–389). Cambridge, MA: MIT Press.

Joos, M. (1948). Language Monograph No. 23: Acoustic Phonetics. *Language*, *24*(2), 5–131.

Kaplan, J. T., Aziz-Zadeh, L., Uddin, L. Q., & Iacoboni, M. (2008). The self across the senses: an fMRI study of self-face and self-voice recognition. *Social Cognitive and Affective Neuroscience*, *3*(3), 218–223. doi:10.1093/scan/nsn014

Knösche, T. R., Lattner, S., Maess, B., Schauer, M., & Friederici, A. D. (2002). Early parallel processing of auditory word and voice information. *NeuroImage*, *17*(3), 1493–1503. doi:10.1006/nimg.2002.1262

Lander, K., & Davies, R. (2008). Does face familiarity influence speechreadability? *Quarterly Journal of Experimental Psychology*, *61*(7), 961–967. doi:10.1080/17470210801908476

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, *26*(4), 708–715. doi:10.3758/BF03211391

Magnuson, J. S., Yamada, R. A., & Nusbaum, H. C. (1995). The effects of familiarity with a voice on speech perception. In *Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan* (pp. 391–392).

Markham, D., & Hazan, V. (2004). The effect of talker- and listener-related factors on intelligibility for a real-word, open-set perception test. *Journal of Speech, Language, and Hearing Research*, *47*(4), 725–737. doi:10.1044/1092-4388(2004/055)

Massaro, D. W. (2004). From multisensory information to talking heads. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 153–176). Cambridge, MA: MIT Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. doi:10.1016/0010-0285(86)90015-0

McGehee, F. (1937). The reliability of the identification of the human voice. *Journal of General Psychology*, *17*(2), 249–271. doi:10.1080/00221309.1937.9917999

McGehee, F. (1944). An experimental study of voice recognition. *The Journal of General Psychology*, *31*(1), 53–65. doi:10.1080/00221309.1944.10545219

Monsen, R. B., & Engebretson, A. M. (1977). Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America*, *62*(4), 981–993. doi:10.1121/1.381593

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234. doi:10.1016/0010-0277(94)90043-4

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. doi:10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. doi:10.1111/j.1467-9280.1994.tb00612.x

Olivos, G. (1967). Response delay, psychophysiologic activation, and recognition of one's own voice. *Psychosomatic Medicine*, *29*(5), 433–440.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328.

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., & Belleville, S. (1994). Functional dissociations following bilateral lesions of auditory cortex. *Brain*, *117*(6), 1283–1301. doi:10.1093/brain/117.6.1283

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184. doi:10.1121/1.1906875

Pollack, I., Pickett, J. M., & Sumby, W. H. (1954). On the identification of speakers by voice. *Journal of the Acoustical Society of America*, *26*(3), 403–406. doi:10.1121/1.1907349

Riding, R. J. (1991). Cognitive styles analysis. In *Learning and Training Technology*. Birmingham.

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects. *Psychological Science*, *18*(5), 392–396. doi:10.1111/j.1467-9280.2007.01911.x

Sanchez, K., Dias, J. W., & Rosenblum, L. D. (2013). Experience with a talker can transfer across modalities to facilitate lipreading. *Attention, Perception & Psychophysics*, 1–7. doi:10.3758/s13414-013-0534-x

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 915–930. doi:10.1037/0278-7393.18.5.915

Schwartz, M. F. (1968). Identification of speaker sex from isolated, voiceless fricatives. *Journal of the Acoustical Society of America*, *43*(5), 1178–1179. doi:10.1121/1.1910954

Schwartz, M. F., & Rine, H. E. (1968). Identification of speaker sex from isolated, whispered vowels. *Journal of the Acoustical Society of America*, *44*(6), 1736–1737. doi:10.1121/1.1911324

Shearme, J. N., & Holmes, J. N. (1959). An experiment concerning the recognition of voices. *Language and Speech*, *2*(3), 123–131. doi:10.1177/002383095900200301

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), 1447–1469. doi:10.1037//0096-1523.28.6.1447

Shipp, T., Qi, Y., Huntley, R., & Hollien, H. (1992). Acoustic and temporal correlates of perceived age. *Journal of Voice*, *6*(3), 211–216. doi:10.1016/S0892-1997(05)80145-6

Shuster, L. I., & Durrant, J. D. (2003). Toward a better understanding of the perception of self-produced speech. *Journal of Communication Disorders*, *36*(1), 1–11. doi:10.1016/S0021-9924(02)00132-6

Studdert-Kennedy, M. (1980). Speech perception. *Language and Speech*, *23*(1), 45–66. doi:10.1177/002383098002300106

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829–834. doi:10.1016/0028-3932(87)90120-5

Van Lancker, D. R., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition*, *1*(2), 185–195. doi:10.1016/0278-2626(82)90016-1

Wood, C. C. (1974). Parallel processing of auditory and phonetic information in speech discrimination. *Perception & Psychophysics*, *15*(3), 501–508. doi:10.3758/BF03199292