THE ROLE OF FACIAL GESTURAL INFORMATION IN SUPPORTING PERCEPTUAL LEARNING OF DEGRADED SPEECH

By

Rachel Victoria Wayne

A thesis submitted to the Department of Psychology

In conformity with the requirements for

the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

September, 2011

ABSTRACT

Everyday speech perception frequently occurs in degraded listening conditions, against a background of noise, interruptions and intermingling voices. Despite these challenges, speech perception is remarkably successful, due in part to perceptual learning. Previous research has demonstrated more rapid perceptual learning of acoustically-degraded speech when listeners are given the opportunity to map the linguistic content of utterances, presented in clear auditory form, onto the degraded auditory utterance. Here, I investigate whether learning is further enhanced by the provision of naturalistic facial gestural information, presented concurrently with either the clear auditory sentence (Experiment I), or with the degraded utterance (Experiment II). Recorded materials were noise-vocoded (4 frequency channels; 50- 8000 Hz). Noise-vocoding (NV) is a popular simulation of speech transduced through a cochlear implant, and 4-channel NV speech is difficult for naïve listeners to understand, but can be learned over several sentences of practice.

In Experiment I, each trial began with an auditory-alone presentation of a degraded stimulus for report (D). In two conditions, this was followed by passive listening to either the clear spoken form and then the degraded form again (condition DCD), or the reverse (DDC); the former format of presentation (DCD) results in more efficient learning (Davis et al, 2005). Condition DC_vD was similar to DCD, except that the clear spoken form was accompanied by facial gestural information (a talking face). The results indicate that presenting clear audiovisual feedback (DC_vD) does not confer any advantage over clear auditory feedback (DCD).

In Experiment II, two groups received a degraded sentence presentation with corresponding facial movements (D_v); the second group also received a second degraded (auditory-alone) presentation (D_vD). Two control conditions and a baseline DC_vD condition

were also tested. Although they never received clear speech feedback, performance in the DvD group was significantly greater than in all others, indicating that perceptual learning mechanisms can capitalize on visual concomitants of speech. The DvD group outperformed the Dv group, suggesting that the second degraded presentation in the DvD condition further facilitates generalization of learning. These findings have important implications for improving comprehension of speech in an unfamiliar accent or following cochlear implantation.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor and mentor, Dr. Ingrid Johnsrude for her generous guidance and support. Her dedication and passion for scientific research is highly infectious and inspires me. I could not have asked for a better and more knowledgeable supervisor and I look forward to continuing to work with her for my Ph.D.!

I am grateful to my Thesis Committee Members: to Dr. Kevin Munhall for having his door open to me as well as for his gracious support and constructive feedback, as well as to Dr. Daryl Wilson for his helpful comments and encouragement. I'd also like to thank my External Examiner, Dr. John Kirby and my Thesis Chair, Dr. Valerie Kuhlmeier.

I am thankful to Julie Buchan, Paul Plante, Dr. Ewen MacDonald and Conor Wild for their assistance with the preparation of the stimuli used in these experiments. I also thank Heather MacDonald for graciously allowing me to film her and for enduring tedious and claustrophobic recording conditions in the name of science.

I am fortunate to work in such a vibrant and cooperative lab environment with such knowledgeable and helpful individuals: Dr. Julia Hyuck, Dr. Fabienne Samson, Conor Wild, Zhuo Zeng, Graham Raynor, Cheryl Hamilton, Kris Marble, Helene Hakyemez, Heather MacDonald and Afiqah Yusuf. I owe a great deal to their assistance and inexhaustible moral support, for completion of this thesis truly would not have been possible without them.

I credit my friends with keeping me sane, particularly Stephanie Allport, Ashley Vesely, and Erin Weinberg for their most helpful advice, their willingness to indulge my incessant and often desultory ramblings, as well as for their unrelenting love and support at even the most unreasonable hour.

I give many thanks to my family, for all of their love and support over the years, particularly my Aunt, Robin Kezwer for her encouragement and helpful advice.

I thank my sister, Brooke Wayne, who continues to inspire me with her kindness and resolve. Her love and support throughout the years has been instrumental. I am also grateful to her for proofreading the final copy of this thesis; few would take on such a painstaking task so willingly!

I am forever indebted to my parents, Martyn and Andrea Wayne, for their fruitful lessons in the value of knowledge and education, and for providing me with a virtually limitless supply of chocolate. I am humbled by their continuous and boundless love, support and encouragement, as well as their extraordinary persistence and determination.

Finally, I acknowledge my Bubbie, Barbara Lieberman, for her *joie de vivre* has helped me to cultivate a strong appreciation for that which truly matters. I owe much to her tenacity and I am grateful to her for being a steadfast beacon of guidance and unconditional support throughout the years. Though she taught me to speak, there are no words to express my love and gratitude for everything that she does.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

GENERAL INTRODUCTION

Speech Perception

The perceptual system for speech is remarkably flexible and readily adapts to multiple variants of speech; much like the perceived color of an object remains constant under varying conditions of illumination, acoustic realizations of the same utterance are perceived to be tokens of the same speech signal. Moreover, speech perception is highly resistant to fluctuations in the signal and everyday speech perception is often challenged by background noise, interruptions and intermingling voices and distortion by room acoustics.

It is only when listening to someone talking very quickly or with a strong accent, or listening in noisy environments or degraded conditions that we become conscious of the robust listening process that the brain often seems to perform with remarkable ease. Although routine speech perception often poses a challenge for normally hearing listeners, these difficulties are considerably exacerbated in the context of hearing impairment. Studying speech perception in noisy and degraded environments may both elucidate the underlying mechanisms of speech perception and inform rehabilitative efforts for hearing loss.

Several factors contribute to the robust nature of speech perception, particularly in noisy or degraded listening conditions. Auditory factors include frequency selectivity (the ability of the auditory system to resolve the different frequencies present in a complex sound) and sensitivity. Other factors involve the use of prior knowledge and context to assist comprehension, use of audiovisual information (the naturalistic lip and facial

movements that accompany speech production), and perceptual learning. The manner in which audiovisual information can support perceptual learning of degraded speech is the focus of this study.

Perceptual Learning of Speech

Adjustments made by the speech system to accommodate for different talkers, speaking rates and accents often occur automatically and rapidly (J. L. Miller, Grosjean, & Lomanto, 1984; J. L. Miller & Liberman, 1979; Smith & Patterson, 2005; Smith, Patterson, Turner, Kawahara, & Irino, 2005; Summerfield, 1981). Alternatively, although heavily accented or degraded speech is difficult to understand upon initial presentation, comprehension often improves upon repeated exposure (Bradlow & Bent, 2008; Clarke & Garrett, 2004; Clopper & Pisoni, 2004; Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Loebach, Pisoni, & Svirsky, 2010; Maye, Aslin, & Tanenhaus, 2008; Weill, 2001). Such improvements in comprehension are evidence for perceptual learning, or "relatively long-lasting changes to an organism's perceptual system that improve its ability to respond to its environment and are caused by this environment" (Goldstone, 1998, p.586). Through perceptual learning, perceptual representations are adjusted in order to assimilate information about the underlying acoustic properties of novel variants of speech. Such retuning can then be used to guide perceptual processes, resulting in subsequent improvement in speech perception.

Support for learning has been documented for a wide range of perceptual processes across modalities (e.g., Ahissar & Hochstein, 2004; Fahle, 2005; Fine & Jacobs, 2002; Fitzgerald & Wright, 2005; Godde, Stauffenberg, Spengler, & Dinse, 2000;

Granger & Lynch, 1991; Hochstein & Ahissar, 2002; Recanzone, Merzenich, Jenkins, Grajski, & Dinse, 1992). This section reviews perceptual learning in conditions or manipulations in which the speech signal has been altered or degraded, outlining the ways in which the perceptual system adapts to the systematic variability present in accented speech, sine-wave speech, time-compressed speech and noise-vocoded (NV) speech.

Research has shown that listeners can quickly adapt to foreign-accented or new dialects of speech, resulting in improved word identification for words previously unheard in that same accent (Bradlow & Bent, 2008; Clarke & Garrett, 2004; Scott & Cutler, 1984; Weill, 2001). For example, Bradlow & Bent (2008) demonstrated significant improvements in recognition accuracy for native English listeners' comprehension of Chinese-accented English after training on either the same talker or multiple Chinese-accented talkers. Clarke & Garrett (2004) measured reaction time for a cross-modal matching task in which participants had to indicate whether a visual probe matched the final word of an English sentence spoken with a Spanish or Chinese accent, as a measure of processing efficiency. The authors showed that listeners' reaction time for foreign-accented speech is initially slower; however, this deficit diminishes within one minute of exposure, indicating a form of perceptual learning.

Alternatively, in time-compressed speech, sine-wave speech, and NV speech, the signal is artificially degraded, either by removing natural spectral or temporal cues present in the speech signal. Time compression of speech produces a signal with fewer pitch periods than the original signal, such that a signal compressed at a 50% compression rate retains only half the number of pitch periods as the original signal. Such a manipulation compresses the signal without deleting portions or creating discontinuities, resulting in a smooth speech signal with preserved spectral characteristics. Since

information is averaged across adjacent pitch periods rather than deleted, the signal retains many of the brief acoustic events, such as release bursts, that are important for phonetic perception. Although sentences of time-compressed speech are initially unintelligible, performance for novel stimuli rapidly improves upon continued exposure, typically reaching a plateau within the first 10-15 sentences (Altmann & Young, 1993; Dupoux & Green, 1997; Mehler et al., 1993; Pallier, Sebastian-Galles, Dupoux, Christophe, & Mehler, 1998).

By contrast, sine-wave speech consists of time-varying sinusoidal patterns that follow the changing formant center frequencies. As most of the traditional speech cues are removed, sine-wave speech sounds unnatural, with unspeechlike timbre and anomalous intonation, much like a set of simultaneous whistles. Using a three-tone time-varying sinusoidal replica of an utterance, Remez, Rubin, Pisoni and Carrell (1981) observed that most listeners did not automatically perceive sine-wave speech as a linguistic entity. However, when instructed that they would hear a sentence, subjects were able to report words from a sine-wave speech utterance, demonstrating speech comprehension in the absence of traditional speech cues (see also, Remez, Nygaard, Rubin, & Howell, 1987). Similarly, Remez, Fellowes and Rubin (1997) created sine-wave utterances from talkers with whom listeners were highly familiar through many years of informal social contact. The authors found that listeners were successfully able to identify their colleagues from the sine-wave sentence presentations, suggesting that sine-wave speech preserves talker-specific aspects of speech. In a follow-up study, Sheffert, Pisoni, Fellowes and Remez (2002) demonstrated that listeners can learn to identify and distinguish different talkers of sine-wave speech, despite removing most of the natural qualities of the speech signal.

Similarly, noise-vocoding is a manipulation that removes much of the spectral information from speech while largely preserving the temporal characteristics (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). As implants currently use at most 22 frequency channels to replace 3,500 hair cells, speech as heard through a cochlear implant is characterized by reduced spectral detail. Learning of NV speech is particularly interesting because it has been used as a simulation of speech as heard through a cochlear implant (an electronic prosthesis that is surgically implanted into the cochlea in order to functionally restore hearing in patients with severe hearing loss). Given that perceptual learning of NV speech is thought to reflect, to some extent, how people with cochlear implants learn to understand transduced speech (Fu & Galvin, 2003; Rosen, Faulkner, & Wilkinson, 1999; Shannon, et al., 1995), an understanding of factors influencing perceptual learning of spectrally-degraded speech may translate into practical strategies for rehabilitating individuals with hearing loss who have been fitted with cochlear implants (e.g., Fu & Galvin, 2007).

The intelligibility of NV speech is easily manipulated through modifying its parameters and the variation and spacing of frequency bands in NV speech is comparable to changing the number and placements of electrodes in a cochlear implant: A greater number of bands or electrodes results in better speech intelligibility (Fishman, Shannon, & Slattery, 1997). Similar to studies of accented speech, sine-wave speech and time-compressed speech, repeated laboratory exposure to NV speech improves listeners' ability to comprehend novel utterances of NV speech, reflecting a form of perceptual learning (Davis, et al., 2005; Hervais-Adelman, et al., 2008; Hervais-Adelman, Davis, Johnsrude, Taylor, & Carlyon, 2011; Loebach & Pisoni, 2008; Loebach, et al., 2010).

Taken together, these accounts of perceptual learning in speech demonstrate the dynamic ability of the human speech perception system to compensate for alterations or distortions present in the speech signal. Rapid perceptual adjustment and improved comprehension of speech with abnormal temporal or spectral properties highlights the flexibility of the perceptual apparatus in adapting to new variants of speech. Crucially, in the examples outlined above, performance generalizes to words never heard before in accented, degraded or altered form. Transfer of learning to novel utterances indicates that listeners are learning something about the underlying regularities present in the signal (e.g., Greenspan, Nusbaum, & Pisoni, 1988), rather than relying on memorization of presented words. The factors that influence the degree and rate of perceptual learning are discussed below.

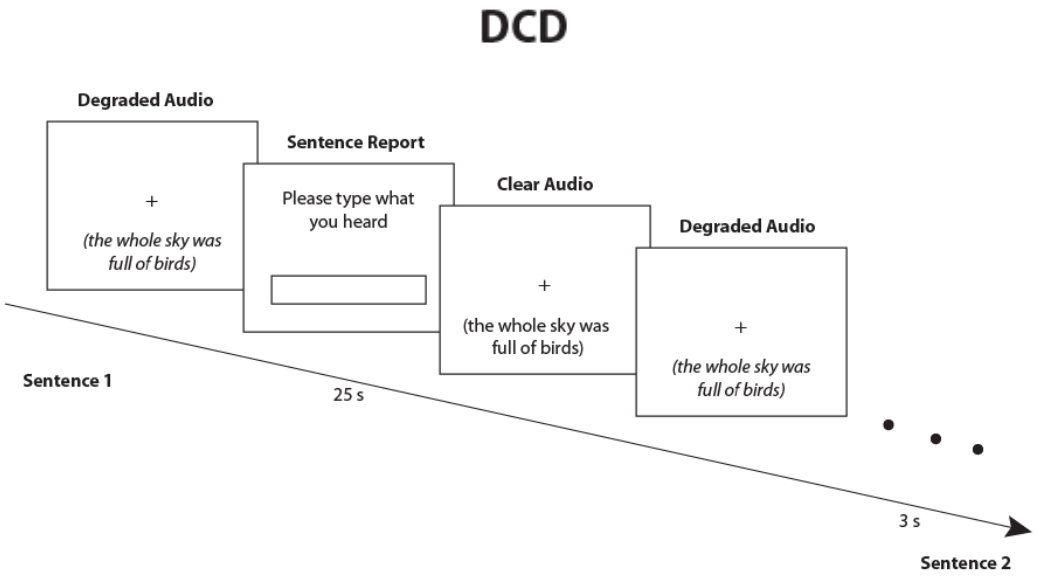
Factors Affecting Perceptual Learning of Speech: The role of Top-Down Processing

A growing body of evidence suggests the presence of both bottom-up and top-down influences on speech processing. Top-down processing refers to the recruitment of higher-level knowledge. For instance, the listener may rely on speaker tone, sentence meaning, sentence structure or sentence context to disambiguate the message content. Top-down processes also include reallocation of resources to increase attention, attending to context, maximizing short-term memory and application of previously acquired knowledge.

Several studies demarcate a contribution for top-down, higher-level linguistic knowledge to perceptual learning of degraded speech and show that learning occurs at a sublexical level, or at a level that permits listeners to learn about speech sounds. Davis et al., (2005) demonstrated perceptual learning for six-band NV speech through repeated

exposure to degraded sentences. In a follow-up experiment, the authors explored the influence of higher-level lexical information on low-level processes using NV speech and found that providing listeners with "feedback" yielded further gains in intelligibility. In this study, subjects were presented with an auditory degraded sentence, and they were asked to report all the words they could understand from the sentence. This word-report score was the dependent measure, taken over successive sentences. Report was then followed by presentation of a clear version of the same sentence (the feedback), followed by a second presentation of the degraded sentence, without report (DCD: distorted-clear-distorted; see Figure 1). Perceptual learning of NV speech (measured as the improvement in word-report scores over trials) was enhanced in the DCD condition, compared to a closely matched condition, differing only in that the clear 'feedback' was presented only after the last degraded presentation (DDC). The authors proposed that feedback may allow listeners to compare the degraded form of the speech utterance with the clear auditory form, enhancing the rate of learning. Hearing the clear version of the utterance rendered the subsequent degraded sentence repetition intelligible to listeners, termed as perceptual 'pop-out' by Davis et al. (2005).

Figure 1. Experimental design for Distorted-Clear-Distorted (DCD) condition in Davis et al. (2005). A degraded sentence was initially presented for report. This was then followed by the clear auditory presentation of the same stimulus, followed by a second degraded sentence presentation. In the Distorted-Distorted-Clear (DDC) control condition, report was followed by a second presentation of the degraded sentence and then the clear auditory version. 30 sentences in total were presented in both conditions.



The authors noted that report scores improved for words that listeners had never heard before in NV form, supporting a role for top-down influences on learning of speech at a sublexical level. Further, the authors demonstrated similar levels of learning when feedback for each vocoded sentence was presented in the form of a written transcription of the degraded sentence. Commensurate learning of NV speech through orthographic feedback implies that improved understanding of NV speech does not require low-level acoustic input. Taken together, the results suggest that learning can be produced through higher-level lexical or phonological information about the content of the sentence and that lexical information about sentence content can modify sublexical processes.

Davis et al., (2005) also showed that sentences made up of real words produced significantly better perceptual learning than non-words. This indication that learning of NV speech is superior for real word sentences was interpreted as evidence for involvement of top-down lexical feedback. However, Hervais-Adelman et al. (2008) reasoned that it was also possible that the sentences composed of non-words could not be retained in phonological working memory and therefore could not be used in the comparison between the clear and degraded speech forms. Accordingly, Hervais-Adelman et al. minimized the load on phonological working memory by using single NV words within a training paradigm. Learning was not significantly different between the real word and non-word conditions, suggesting that learning may involve sublexical phonological rather than lexical units. Thus, the evidence suggests that although top-down lexical information may support perceptual learning, it is not required for perceptual learning of NV speech.

Further evidence for top-down influences on perceptual learning occurring at a sublexical level comes from studies of accents, ambiguous fricatives and time-compressed speech. Maye, Aslin and Tanenhaus (2008) created a novel, synthetic accent of English by systematically lowering front vowels (e.g., 'wetch' instead of 'witch'). A 20-minute story was read out to participants at two different time points, produced either in standard American English (Day 1) or in accented English (Day 2). On both days, after hearing the story, participants performed an auditory lexical-decision task. Items in the lexical-decision task had either standard American English front vowels or accented (lowered) English front vowels and half these items occurred in the stories, whereas the other half were new. Participants more often and more quickly indicated that items with lowered front vowels ('wetch') were words after hearing the story read in accented

English than after hearing it in Standard English. Moreover, this increased endorsement of accented English items on Day 2 was significant even for new items, indicating that participants learned something about the general phonetic characteristics of the accent (i.e., that front vowels are lowered), suggesting a sublexical locus of learning. In a follow-up experiment, the authors demonstrated that listening to the accented stories results in a change that is specific to the direction of the accent, rather than simply allowing for more noise or a general broadening of front vowel categories. These results demonstrate that the perceptual apparatus can make systematic changes in the mapping of acoustic-phonetic input to lexical representations. The authors also show that learning is dependent upon the stimuli to which listeners are exposed, although the results do not allow for more specific conclusions.

A study by Norris, McQueen and Cutler (2003) yielded insight into the conditions required for perceptual learning involving phonemic category boundaries. The authors used a lexical decision task to influence Dutch listeners' interpretation of ambiguous fricatives at the end of a word. These lexically influenced interpretations then influenced subjects' interpretation of ambiguous sounds in a subsequent fricative phonemic categorization task. Listeners who heard the ambiguous sound in the context of /f/-final words subsequently categorized more items on an /f/-/s/ continuum as /f/, whereas listeners who heard the same ambiguous sound in /s/ final words categorized more items on the same /f/-/s/ continuum as /s/. Crucially, the authors also demonstrated that lexical contexts were required for perceptual learning in altering phonemic category boundaries; no shift was observed for listeners hearing the ambiguous sound in non-words.

Eisner and McQueen (2005) also demonstrated that perceptual learning of fricative sounds occurs at a sublexical level. Using the same paradigm as Norris et al.

(2003), the authors demonstrated a shift in categorization boundaries when the test fricatives appeared after vowels spoken by novel talkers of either the same or opposite gender, but not if both the vowel and the fricative were produced by a novel talker. The authors also demonstrated that a shift in phonemic category boundaries for fricatives of a novel talker's speech occurred only when the fricative of the novel talker had been spliced into the exposure talker's speech during training. Since evidence for a shift in categorization boundaries generalized to new speakers (who produced the vowel sound before the critical phoneme), the authors concluded that learning occurs at a phonemic level. Thus, while listeners appear to use lexical contexts to inform their interpretation of an ambiguous fricative (in Norris et al., 2003), categorization of ambiguous fricatives appears to be talker-specific.

Further evidence for perceptual learning on a sublexical level was provided by Kraljic and Samuel (2006). Similar to Norris et al. (2003), Kraljic and Samuel demonstrated that perceptual learning for a stop consonant generalized to a new consonant that had the same featural voicing relationship. This confirmed that listeners were broadly retuning phonetic categories, rather than modifying more abstract levels of representation. The results of these studies demonstrating shifts in phonetic category boundaries suggest that information from the lexicon can modify prelexical processing, demonstrating that listeners can use higher-level information to tune lower-level processes and thus constrain interpretation of the ambiguous fricative (see also Samuel & Kraljic, 2009).

Two other studies of time-compressed speech also suggest the involvement of sublexical units in the perceptual learning of speech. Altman and Young (1993) demonstrated improvement in comprehension of compressed sentences after listening to

nonsense utterances of time-compressed speech. Such results suggest that lexical level speech recognition is not required for learning of time-compressed speech. Pallier and colleagues (1998) exploited the fact that the Catalan and Spanish languages both share many phonological properties, relying on similar speech perception routines (Sebastian-Gallés, Dupoux, Segui, & Mehler, 1992). The authors found that bilingual speakers of Catalan and Spanish who were trained on time-compressed speech in one language showed significant improvements in performance when tested with time-compressed speech in the second language compared to baseline. More importantly, the study also demonstrated transfer of learning from one language to the other in monolingual speakers of either Catalan or Spanish. There was also some transfer of learning from English to Dutch sentences of time-compressed speech for monolingual English speakers (English and Dutch share some supra-segmental phonological properties in that they are stress-timed languages, although there is less overlap between them than between Catalan and Spanish). Such transfer of learning was not observed in bilingual speakers of French and English (which are phonologically more distinct from one another than are English and Dutch). Similarly, Bent et al. (2011) showed that native-English speakers trained on German sine-wave-vocoded sentences performed comparably on English sine-wave-vocoded test sentences to those trained and tested on English sentences. These results suggest that transfer of learning of degraded speech between languages occurs only when both languages contain shared supra-segmental phonological properties, even when listeners are fluent in only one of the languages. Since comprehension was not required for learning, this suggests the involvement of sublexical processes.

In summary, the perceptual learning literature for speech implicates a sublexical locus of learning in which higher-level information may help to disambiguate the lexical

intent of a spoken utterance. Davis et al. (2005) and Hervais-Adelman et al. (2008) suggested that sublexical (or lexical) information in the sentence content helps to maintain a phonological representation of the target item in short-term memory for comparison with the degraded utterance. Such a comparison between the degraded input and the phonological representation may help to tune sublexical auditory processes so that subsequently presented degraded utterances can be perceived more clearly. Higher-order mechanisms may thus be crucial for perceptual learning by helping to constrain interpretations of degraded (or ambiguous or time-compressed) speech in order to reinforce more accurate perceptual hypotheses. Bradlow and Bent (2008) proposed a similar explanation after observing faster learning for more intelligible than for less intelligible foreign-accented speech, positing that better learning in the high-intelligibility condition was facilitated by more consistent and accurate feedback from higher linguistic levels. Can these perceptual hypotheses about the content of a degraded utterance be further constrained? An additional source of information that the perceptual system may be able to exploit in learning is the focus of the next section.

Audiovisual Integration in Speech

Although the studies discussed above have examined speech as relayed through auditory channels, listening to speech in the real world is rarely a unimodal process; everyday speech has both seen and heard correlates and there is a wealth of research on the complementary nature of audio and visual information in speech. It has been shown that visual cues can improve the signal-to-noise ratio (SNR) by up to 15 dB (Sumby & Pollack, 1954) and a 1 dB increase in SNR can correspond to improvements of 5-10% in intelligibility (Grant & Braida, 1991; G. A. Miller, Heise, & Lichten, 1951). Further, it is

known that speech intelligibility is enhanced by providing visual information about facial and head movements (Grant & Seitz, 2000; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Visual cues are beneficial even in ideal listening conditions (Arnold & Hill, 2001; Reisberg, McLean, & Goldfield, 1987; Remez, 2005) and visual speech enhances the intelligibility of auditory speech when it is heavily accented or presented in background noise (MacLeod & Summerfield, 1990; Reisberg, et al., 1987). In many languages, the auditory and visual components of speech are complementary such that segment distinctions that are harder to hear are easier to see, and vice versa (Rosenblum, 2008).

Audiovisual integration is a natural and ubiquitous phenomenon, and a strong case for the automaticity of multimodal speech integration has been made through studies examining the widely cited McGurk effect, in which incongruent audiovisual information can change the identity of a speech percept. McGurk and MacDonald (1976) demonstrated illusory perception of speech sounds not presented in either modality alone. The authors used presented combinations of auditory and visual concomitants of syllables and words: these were presented to appear coherent, but were physically impossible (i.e., the auditory syllable 'ba', which requires lip closure, heard while the visual face spoke 'ga' with no closure). The resultant percepts are perceived to be somewhere midway between the two segments (for example, a auditory /ba/ dubbed onto a visual /ga/ is heard as 'da'). The McGurk effect persists even when subjects are explicitly told about the dubbing procedure or are asked to attend only to information in one modality (Massaro, 1987). Moreover, the two information streams do not need to come from the same source; integration can occur for auditory and visual inputs generated by speakers of different genders (Green, Kuhl, Meltzoff, & Stevens, 1991).

There is also evidence for cross-modal benefits in audiovisual integration, exemplifying the bidirectional interaction between auditory and visual speech information. Rosenblum, Miller and Sanchez (2007) demonstrated that subjects who lipread from a talker and then heard that talker in a speech-in-noise task performed better on the latter task than subjects who lipread from one talker then heard another talker. Auditory speech has also been shown to facilitate lipreading when the lipread stimulus is ambiguous (Baart & Vroomen, 2010). Visual information appears to "speed up" neural processing of auditory speech, resulting in an amplitude reduction of auditory evoked potentials (Besle, Fort, Delpuech, & Giard, 2004; van Wassenhove, Grant, & Poeppel, 2005).

The results of a number of neuroimaging studies have been interpreted as evidence for early audiovisual integration, suggesting that integration may occur as early as the auditory cortex (Callan, Callan, Kroos, & Vatikiotis-Bateson, 2001; Calvert et al., 1997; Colin et al., 2002; Pekkola et al., 2005; Sams et al., 1991). This is a stage of processing before segments are phonetically categorized, possibly even before phonetic features are extracted (Green, 1998; Green & Gerdeman, 1995; Green & Miller, 1985; Green & Norrix, 2001). Further, there is evidence that a silent lipreading task activates the primary auditory cortex in a manner similar to heard speech (Calvert, et al., 1997; MacSweeney et al., 2000; MacSweeney et al., 2002), although this has been challenged (Bernstein et al., 2002). Putative evidence for early integration of audiovisual information in speech raises the possibility that perceptual representations of speech may be either bimodal or modality neutral.

Correlated time-varying dimensions of auditory and visual signals may provide a basis for audiovisual integration. Summerfield (1987) suggested that auditory and visual

inputs might be evaluated along a common metric, based on the kinematics and dynamics of underlying articulatory behavior (Bernstein, Auer, & Moore, 2004; Liberman & Mattingly, 1985; Rosenblum, 1994; Schwartz, Robert-Ribes, & Escudier, 1998). Similarly, in their work on cross-modality comodulation, Grant and Seitz (2000) observed that visual cues derived from dynamic facial movements during speech production interact with time-aligned auditory cues to enhance sensitivity in auditory detection, with the degree of visual influence depending on the correlation between acoustic envelopes and visible articulator movements. Lachs and Pisoni (2004) suggested that cross-modal information may be specified in dynamic temporal information; the authors demonstrated that cross-modal matching of auditory and visual stimuli is only possible when transformations preserve spectral and temporal patterns of formant frequencies as they change over time. Similarly, Lander, Hill, Kamachi, and Vatikiotis-Bateson (2007), suggested that speech rhythm and expressivity provides a bimodal, dynamic identity signature (see also Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003). Taken together, these studies highlight the importance of the time-varying dimensions of the auditory and visual signals in providing basis for audiovisual integration, possibly through amodal speech representations.

Contributions of Articulatory Representations to Learning

Given the close association between auditory and visual aspects of speech, it is perhaps unsurprising that motor regions of the brain may support speech perception (Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007; Pulvermuller et al., 2006; Watkins, Strafella, & Paus, 2003; Wilson, Saygin, Sereno, & Iacoboni, 2004). It is important to note that many of these

neuroimaging studies are correlative in nature, and only allow for speculation as to the directionality of causation.

The evidence suggests that motor regions are recruited to a greater extent or preferentially when speech is presented in noise (Callan et al., 2003; D'Ausilio, Bufalari, Salmas, & Fadiga, 2011) and that motor regions may aid comprehension of speech that is difficult to understand. Callan et al. (2003) also found that these regions are recruited especially when high quality visual information is available. Skipper et al. (2005) showed that the visual aspects of observable articulatory gestures (rather than the auditory content of speech) is primarily responsible for the activation of motor regions in audiovisual speech perception, outlining a contribution for visual information in activating motor regions. Such evidence for pre-motor contributions nicely complements the modality-neutral account for speech perception that has been advanced within the audiovisual speech perception literature, outlining a mechanism for their common representation via articulatory templates.

Research by Hervais-Adelman, Carlyon, Johnsrude and Davis (submitted) suggests that motor representations may be involved in perceptual learning of NV speech. The authors postulated that mapping of clear speech onto the degraded speech form may involve non-acoustic articulatory representations. The authors compared imaging data for subjects listening to clear speech versus potentially intelligible NV words. They found that in addition to activating areas associated with listening to clear speech, listening to NV speech also activated the regions extending from precentral gyrus into left frontal operculum, implicating involvement of pre-motor and prefrontal regions, respectively. Moreover, activation in the precentral gyrus was located close to areas implicated in speech production, namely to the site of mouth and tongue representations (Pulvermuller,

et al., 2006; Wilson, et al., 2004). TMS studies have also demonstrated increased excitability of the speech production system during auditory speech perception (Fadiga, et al., 2002; Watkins, et al., 2003), which appears to be modulated by activity in Broca's area (Watkins & Paus, 2004). The involvement of these motor regions in speech perception lends credence to the idea that these regions are involved in decoding speech using non-acoustic, articulatory templates (Davis & Johnsrude, 2007; Iacoboni, 2008; Poeppel & Monahan, 2010). Such regions may provide an internal simulation of articulatory gestures that helps to match degraded speech input onto internal templates derived from a prototypical motor pattern.

Study Rationale

The combined literature on perceptual learning, audiovisual integration and recruitment of motor areas in speech delineate a contribution for facial gestural information to perceptual learning of NV speech. Non-acoustic articulatory templates might facilitate top-down support of speech comprehension. These might provide a mechanism for a discrepancy comparison between the clear and degraded speech forms, matched through the time-varying properties of the speech signal in the auditory and visual modalities. The provision of visual speech cues as feedback for degraded speech might further facilitate this template-matching process, possibly through greater recruitment of motor templates. Facial gestural information could be used to further disambiguate the degraded speech, which would provide greater constraint on perceptual hypotheses. This would in turn provide more accurate mapping of degraded speech onto phonological representations, resulting in greater gains in learning compared to unimodal auditory feedback.

In this study, I expand upon the work of Davis et al. (2005), evaluating the contribution of visual speech cues to perceptual learning of NV speech. In Experiment I, I examine the rate of learning in an audiovisual clear feedback condition, compared to the auditory-alone clear feedback condition used by Davis et al. In the second experiment, I evaluate learning within a more naturalistic paradigm, in which the degraded auditory input is presented simultaneously with clear visual speech, in order to better simulate the experience of CI users. I also investigate the relative contribution of a second degraded sentence presentation to learning as a subsidiary aim.

Applications

It is estimated that 25% of individuals aged 65-75 years and 70-80% of individuals over the age of 75 suffer from age-related sensorineural hearing impairment (Medline MedlinePlus, 2010), and this number is expected to rise (World Health WHO, 2001). Correlated with a decreased quality of life (World Health WHO, 2002), age-related hearing loss is a leading cause of a higher score on "years lived with disability" scales in the adult years (Cohen, Labadie, & Haynes, 2005; LaForge, Spector, & Sternberg, 1992; Mulrow et al., 1990) and contributes to social isolation (Mulrow, et al., 1990) and depression (Carabellese et al., 1993). Age-related auditory deficits are also associated with cognitive decline (Arlinger, 2003; Schneider, Daneman, & Pichora-Fuller, 2002). The evidence suggests that cognitive deficits are exaggerated when auditory input is compromised (Uhlmann, Larson, Rees, Koepsell, & Duckert, 1989), and almost half of mildly hearing impaired patients with dementia improve when hearing is restored (Allen et al., 2003). Taken together, the evidence demonstrates that rehabilitation

of hearing loss is likely to provide both functional and psychological contributions to quality of life.

Research shows that older listeners (aged 63-75) require a 3 dB more favourable signal-to-noise ratio (SNR) than their younger counterparts (Li, Daneman, Qi, & Schneider, 2004; Pichora-Fuller & Singh, 2006). Although hearing assistive technologies are capable of providing functional gain in many patients, gains in speech recognition are insufficient for the estimated 10% of the older adult population with more severe to profound hearing losses (thresholds greater than 70 dB) above 1 kHz (Ching, Dillon, & Byrne, 1998). Cochlear implants are a popular (NICD, 2009), efficacious (Kelsall, Shallop, & Burnelli, 1995; Labadie, Carrasco, Gilmer, & Pillsbury, 2000; Shin et al., 2000) and cost-effective (Francis, Chee, Yeagle, Cheng, & Niparko, 2002) solution for older adults coping with severe hearing loss. Use of cochlear implants in older adults has also been shown to result in improvements in overall well-being (Francis, et al., 2002; Horn et al., 1991).

However, speech as heard through a cochlear implant is very different from normal, acoustic speech. Months or even years of auditory rehabilitation therapy is advised for users (Carlsson, Hall, Lind, & Danermark, 2011; Francis, et al., 2002). With training, sentence recognition performance may reach an impressive 72.5% for post-lingually deafened older adults at 12 months post-operative evaluation (Pasanisi et al., 2003), although there is dramatic intersubject variability in asymptotic performance levels.

Yet, prelingually deafened cochlear implant recipients rarely undergo any long-term systematic rehabilitation to improve speech perception (Clark, 2003; McConkey-Robbins, 2000) and there is currently little consensus regarding best rehabilitative

practices (Boothroyd, 2007). Thus, development and implementation of empirically-derived strategies for rehabilitation of newly implanted cochlear implant users that are optimally efficient and cost-effective is essential, given the growing population of individuals with hearing loss.

CHAPTER 2

EXPERIMENT I: THE ROLE OF AUDIOVISUAL FEEDBACK IN SUPPORTING PERCEPTUAL LEARNING OF NOISE-VOCODED SPEECH

Experiment I Introduction

As described above, Davis et al., (2005) demonstrated more rapid learning of NV speech for listeners who were given the opportunity to map the degraded speech utterance onto the clear auditory speech form. This experiment evaluates the relative contribution of facial gestural information presented concurrently with the clear auditory speech form. If individuals can productively use facial gestures to help tune perception, as they appear to use lexical or phonological information, then I should find increased learning in this group compared to an auditory-only-feedback group.

Method

Participants

Thirty-six participants (5 males, 31 females) from Queen's University were tested. All were native English speakers aged between 17-28 years with no history of hearing or language impairment. Twenty-eight participants were right-handed and two participants were left-handed (handedness data was not available for six participants). All participants had normal or corrected-to-normal vision, had no known verbal learning disability and no previous exposure to NV speech. Subjects received either course credit or monetary compensation for their participation. The study was cleared by Queen's General Research Ethics Board.

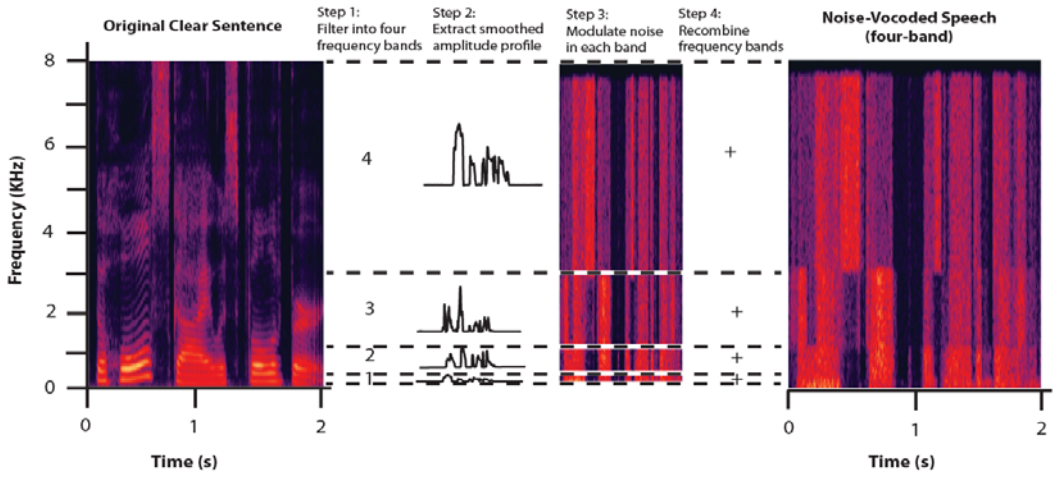
Design and Materials

Materials for this experiment consisted of six sets of five vocoded sentences (designated A, B, C, D, E and F), for a total of 30 sentences. I used simple, declarative sentences with a range of lengths (6 to 13 words, *M* = 9 words/sentence, see Appendix A). Sentence materials were essentially those of Davis et al (2005), with slight wording alterations (e.g., "lorry" changed to "truck') for our Canadian participants. Sentences were assigned to groups such that they were matched item-by-item for number of words, spoken duration, and naturalness and imageability (as rated in a pilot study; see Rodd, Davis, & Johnsrude, 2005). Audiovisual materials were recorded from a young, female native speaker of North American English in a sound-attenuating booth. The video was recorded through a Panasonic AG-DVC7 video camera, and sound was recorded through the video camera as well as through an AKG C1000S microphone into an RME Fireface 400 audio interface (sampling @ 16-bits, 44.1 khz) connected to a PC running Adobe Audition.

The edited soundfile for each of the 30 sentences was processed to create a four-band NV version (see Figure 2). NV stimuli were created as described by Shannon et al. (1995) using a custom Matlab vocoder. Items were filtered into four contiguous frequency bands (cutoffs: 50Hz → 369Hz → 1160Hz →3124Hz →8000Hz); band cutoffs were selected to be equally spaced along the basilar membrane (Greenwood, 1990) and were implemented using finite impulse response (FIR) Hann band-pass filters (window length of 801 samples). The amplitude envelope from each frequency band was extracted by full-wave rectifying the band-limited signal, and then low-pass filtering it at 30Hz using a fourth-order Butterworth filter. The resulting envelopes were then applied to band-passed noise in the same frequency ranges. Finally, the four bands of modulated noise were

added to produce the NV utterance. After processing, all stimuli (clear and NV speech) were normalized to have the same average RMS power.

Figure 2. Processing steps involved in transforming clear speech into noise-vocoded speech (left spectrogram). Sentences are filtered into four non-overlapping frequency ranges (Step 1), the amplitude envelope in each band is extracted and smoothed (Step 2), and wide-band noise in each frequency range is modulated using this amplitude envelope (Step 3) and combined to produce a noise-vocoded sentence (Step 4; right spectrogram) (after Davis et al., 2005).



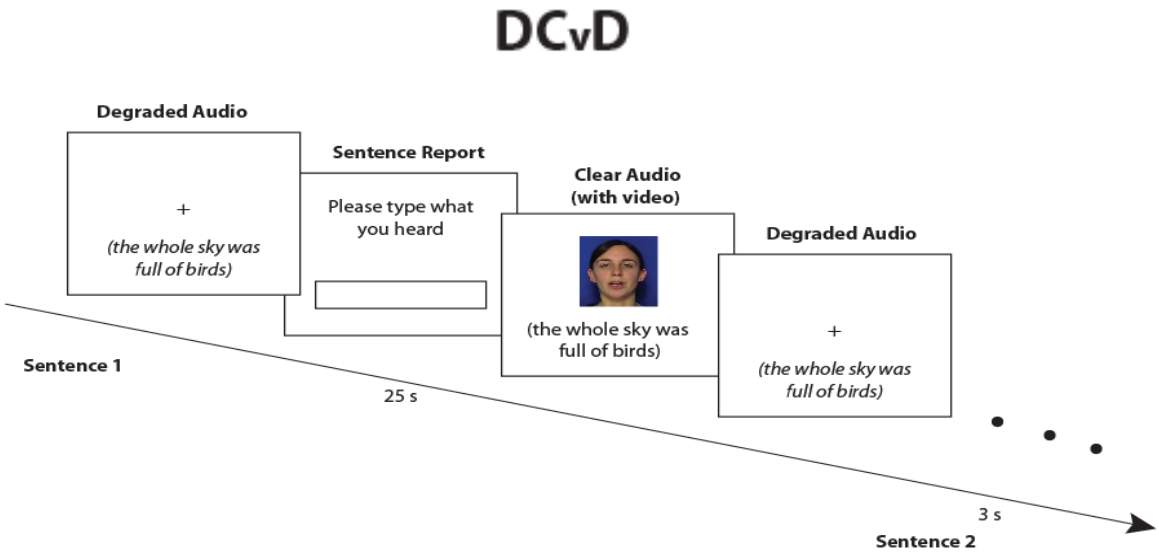
The auditory component of the videocamera file was extracted using Virtual Dub (a free downloadable video capture and processing utility) and the waveform on screen was used as a template to align the normalized clear Adobe Audition waveform in time. The normalized clear audio file was then recombined with the video file using Virtual Dub in order to produce the clear audiovisual stimuli.

Procedure

Participants were randomly assigned to one of 3 conditions, for a total of 12 subjects in each condition. For all conditions, each trial began with an initial presentation

of the vocoded sentence ('D'), followed by the participant reporting all the words they could understand from the sentence. Participants were then presented with two repetitions of each sentence, the precise nature of which depended on condition. In the DDC condition, they heard the sentence vocoded again, and then clear (DDC; degraded audio-degraded audio-clear audio). In the DCD condition, they heard the clear version of the sentence, followed by the vocoded version again (degraded audio-clear audio-degraded audio), as in Davis et al. (2005). The DCvD condition was very similar to the DCD condition, except that the clear speech was accompanied by a video displaying the corresponding facial gestures in the form of a talking face (DCvD; degraded audio-clear audiovisual-degraded audio; see Figure 3). Report was always taken before subjects were presented with information about the content of the sentence.

Figure 3. Format of presentation in the DCvD condition. The condition is similar to the DCD condition except that clear audiovisual feedback is presented (rather than auditory-alone feedback) following the initial degraded sentence presentation.



All subjects received all six sentence sets (tested at 6 time points), for a total of 30 sentences. Sentences within each set were presented in a fixed order to all participants, as in Davis et al, (2005). However, the order of the sentence sets was counterbalanced across conditions. All sentence sets were presented an equal number of times in each condition, at each time point. Such counterbalancing allows us to compare performance on the blocks of sentences without any confound produced by differences in the difficulty of various sentences (Pollatsek & Well, 1995).

Participants were fitted with Sennheiser HD 265 headphones and were tested in a single-walled sound-attenuating booth (Eckel Industries), guided by a computer running Eprime. All sentences were played at 65-70 dB SPL(A). Subjects were instructed to listen carefully and to report as many words as they could from the initial presentation of each vocoded sentence, typing directly into Eprime. Subjects were given 25 seconds for report. In order to ascertain that subjects' working memory capacity was sufficient to perform the task, subjects were prompted to type out four sentences (6 to 10 words, $M = 8$ words/sentence, see Appendix B), presented as clear speech prior to the start of testing (all participants obtained perfect scores on the memory-test sentences, indicating that working memory capacity was not a limiting factor in their report scores). Subjects then heard a single sample sentence vocoded using 15 bands. This is easily intelligible even to naïve listeners, but served as an example of the form of distortion that they would hear in the task. A secondary fifteen minute speech-in-noise task was administered to all subjects prior to beginning the experiment. Preliminary analysis indicates that the findings are unrelated to the results of this experiment. Thus, this task and the results will not be reported here.

Data Coding and Analysis

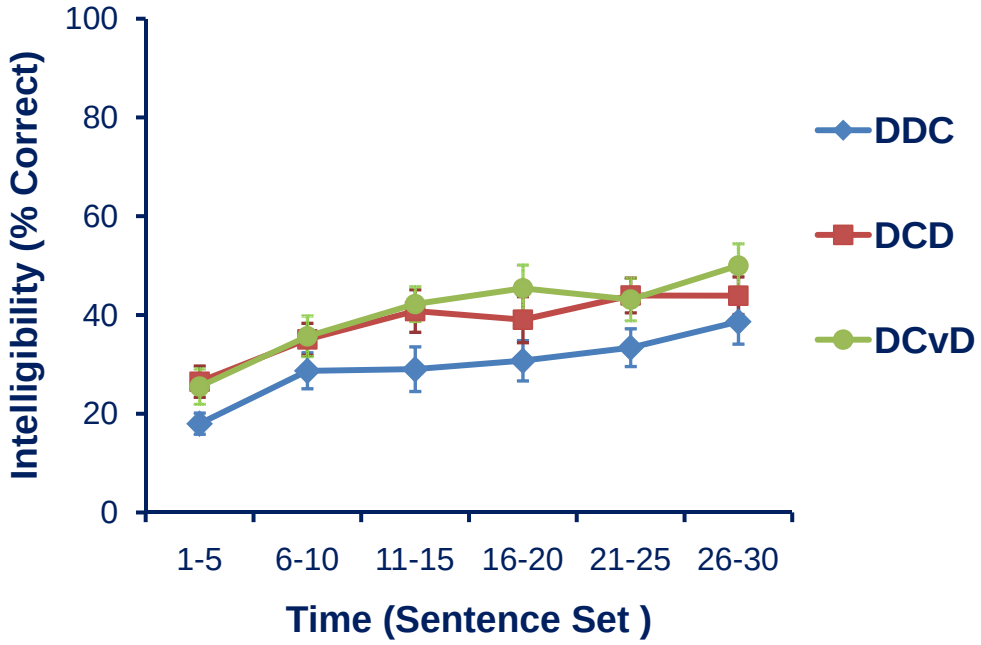
Participants' written reports of the clear sentences in the memory test and vocoded sentences in the main experiment were scored for the percentage of words in each sentence that were reported correctly. As in Davis et al. (2005), words were scored as correct only if there was a perfect match between the written form and the word produced in the sentence (morphological variants were scored as incorrect, whereas homonyms and misspellings, even if semantically anomalous, were scored as correct). Words were not scored as correct if they were reported in the wrong order, but words in the correct order were scored as correct even if intervening words were absent or incorrectly reported.

I conducted analyses both by participants and by items in order to confirm that any change in report scores over time and between conditions did not result from differences in sentence difficulty. Accordingly, I conducted 2-factor ANOVAs with Condition (3 levels) and Time (6 levels) on report scores, averaged over participants or items. Analysis by participants allows us to examine performance across conditions, but variability across sentence materials may occlude significant effects if large enough. Analysis by items allows us to compare performance on the same sentence across different conditions and at different time points, but variability across subjects may occlude significant effects if large enough. These two analyses, taken together, compensate for both item and subject variability. An additional dummy variable in the items analysis codes for which group of participants was tested first on each group of sentences, although main effects and interactions involving this dummy variable will not be reported (Pollatsek & Well, 1995). Condition was entered as a between-subjects factor in the analysis by participants and as a within-subjects factor in the analysis by items.

Results

Analysis by participants (F_1) and items (F_2) indicated a main effect of time; report scores improved over the 6 sets of sentences, $F_1(5, 29) = 20.821$, $p < .001$, $\eta_p^2 = .782$; $F_2(5, 20) = 8.504$, $\eta_p^2 = .680$, $p < .001$, similar to Davis et al. (2005) (Figure 4). There was a main effect of condition that was reliable both by participants and items, $F_1(2, 33) = 3.785$, $MS = .225$, $p = .033$, $\eta_p^2 = .187$; $F_2(2, 23) = 28.273$, $p < .001$, $\eta_p^2 = .711$. Pairwise comparisons revealed that subjects' performance in the DCD condition (M = 38.2%, SD = 11.4%, range = 21.4- 55.0%) was significantly better than in the DDC condition (M = 29.8%, SD= 12.2%, range = 15.6- 48.3%, $p_1 = .045$; $p_2 < .001$), replicating Davis et al. (2005). Similarly, performance for subjects in the DCvD condition (M = 40.3%, SD = 11.5%, range = 27.3- 66.8%) was significantly better than for subjects in the DDC condition ($p_1 = .014$; $p_2 < .001$). Contrary to our hypothesis, performance was not significantly different between the DCvD and DCD conditions either by participants or items; presenting clear feedback in audiovisual format did not confer any advantage over an auditory-alone format. The Time x Condition interaction was nonsignificant both by participants and items.

Figure 4. Experiment I average report scores. Report scores from Experiment I averaged over six groups of five sentences in three groups of participants. Error bars show plus or minus one standard error of the mean over participants.



Discussion

The results replicate the findings of Experiment II of Davis et al. (2005) and are consistent with the idea that gains in learning are conferred by an opportunity to map the clear version of the sentence onto the degraded speech form. However, the finding that performance did not significantly differ between the DCD and DCvD conditions indicates that providing accompanying facial movements alongside clear auditory feedback yields no additional benefit over presentation of auditory-alone feedback. This may be related to the redundancy of the visual information in the DCvD (audiovisual feedback) condition: Since the clear auditory feedback is fully comprehensible in the DCvD condition, it is quite possible that the visual component offers very little, if any, gain to listeners.

However, the clear audiovisual and degraded speech forms were presented sequentially, whereas the perceptual apparatus is designed to capitalize on synchronous audiovisual information. A more naturalistic condition is the focus of Experiment II.

CHAPTER 3

EXPERIMENT II: THE ROLE OF CONCURRENT FACIAL GESTURAL INFORMATION IN SUPPORTING PERCEPTUAL LEARNING OF NOISE-VOCODED SPEECH

Introduction

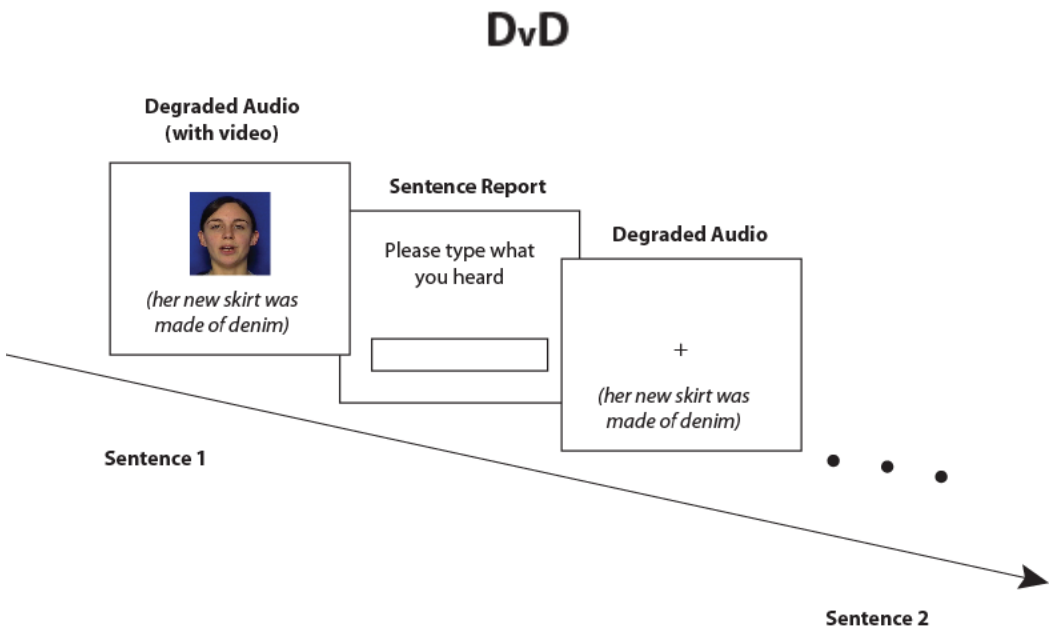
The speech signal is redundant in that speech information is represented in both the auditory and visual modalities. The complementary nature of synchronous auditory and visual speech information can thus be exploited by the perceptual system, particularly when the fidelity of the signal in one modality is compromised. This study explored whether listeners can exploit facial gestural information presented in the visual modality in order to improve perception of degraded acoustic information.

I test whether presenting facial gestural information concurrently with the degraded utterance improves the rate of perceptual learning. Through control conditions, I also test whether pop-out may be driving learning (as suggested by Davis et al., 2005). I hypothesized that learning would be more efficient when participants have the opportunity to directly map visual facial gestural information onto the degraded speech form. Presentation of corresponding facial gestural movements simultaneously with the degraded sentence may best enable the perceptual system to capitalize on supplemental visual information to help identify the degraded speech form. This concordance between audio and visual information might further constrain appropriate perceptual hypotheses (Hervais-Adelman et al., 2008), thereby improving speech comprehension for subsequent degraded sentences.

In Experiment II, I test two naturalistic conditions that more closely approximate the listening situation of listeners using a cochlear implant. In the first condition, the

degraded speech form was presented with the corresponding facial gestures (D_v). The second condition was identical to the D_v condition, except that the degraded auditory speech form was presented a second time (without synchronous facial gestural information; D_vD; see Figure 5). This condition was intended to test whether a second presentation of the degraded sentence would further reinforce learning.

Figure 5. D_vD format presentation. Subjects first heard the degraded sentence presented with concurrent facial gestural information. Following report, subjects were provided with an auditory-only repetition of the same degraded sentence.



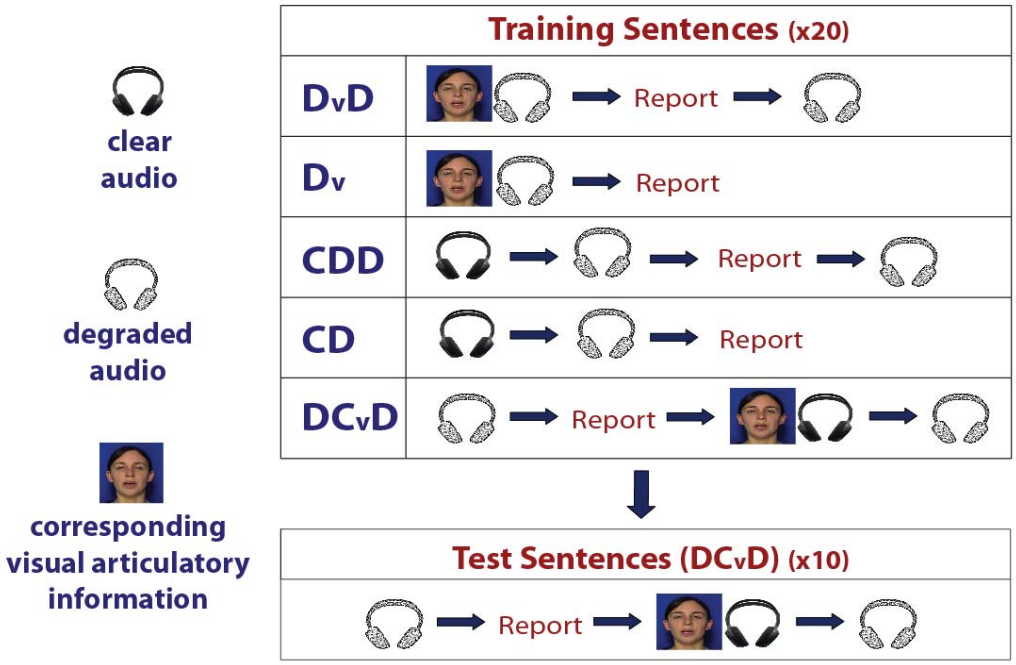
I chose to use the DC_vD condition from Experiment I as a baseline in order to investigate whether simultaneous facial gestural 'feedback' for degraded sentences confers greater benefit than clear audiovisual feedback presented after a degraded audio-only presentation.

I reasoned that any greater benefit observed in the more naturalistic D_v or D_vD condition compared to DC_vD would suggest one of two possibilities. First, gains in perceptual learning of degraded speech may be attributable to the presence of synchronous audiovisual information. However, it is also possible that benefits may be derived from the greater overall intelligibility of the D_v stimulus in the D_v and D_vD conditions, compared to an auditory-only degraded stimulus. Thus, I also tested two control conditions in which listeners were supplied with complete linguistic information prior to hearing the first degraded sentence presentation.

In the first control condition, CD, subjects heard the degraded speech form knowing the identity of the sentence; subjects first heard the clear auditory speech form, followed by a degraded auditory presentation of the same utterance (see Figure 6). As D_vD is the only experimental condition in which there are two 'usable' degraded sentence presentations (i.e., the only condition in which there is sufficient information to render the two degraded presentations reasonably intelligible), it is important to examine whether any superior performance in D_vD is due to greater exposure to NV speech. Therefore, in control condition CDD there are two degraded sentence presentations following a single presentation of the clear auditory speech form. In addition, the phenomenon of pop-out as outlined by Davis et al. (2005) strongly suggests that inducing perceptual pop-out for degraded speech drives learning and that the magnitude of pop-out is related to the completeness of linguistic information when the acoustic input is degraded. Therefore, these two controls also allow us to test whether conditions that are believed to induce perceptual pop-out promote learning; since both the CDD and CD conditions will produce very strong pop-out (as per Davis et al., 2005), superior performance for subjects in the

DvD condition relative to these conditions would imply that facilitated learning is not completely dependent upon pop-out but must be driven by other factors.

Figure 6. Training and test format for the five conditions in Experiment II. Subjects were trained on 20 sentences presented in a format according to their assigned condition. All subjects were then tested on ten counterbalanced sentences presented in a common DCvD format.



In Experiment I, I measured learning as change in word report score to the first presented degraded stimulus ('D') over trials. This is obviously confounded in Dv, DvD, CD and CDD conditions since the participant has additional information at the time of the first 'D' presentation on each trial that very likely renders these sentence presentations more intelligible than degraded auditory-only utterances. Thus, in order to compare

learning across conditions, I evaluated how well listeners in the different conditions could generalize their learning to a common set of auditory-only degraded sentences.

Following the training phase (with sentences presented in either D_v, D_vD, CD and CDD format), ten sentences were presented in a common DC_vD format (the test phase), to all subjects (see Figure 6). Word report scores after the initial degraded presentation on each DC_vD trial could be compared among the groups trained on the different conditions. A fifth, naïve control group did not receive any training, and was tested on the same DC_vD sentences. As mentioned above, such a baseline condition will allow us to compare the performance of subjects presented with concurrent 'feedback' in the D_v and D_vD conditions with the performance of subjects presented with clear auditory feedback (as in Experiment I and in Davis et al., 2005).

Methods

Participants

Ninety participants (58 females, 18 males; the sex of the remaining 14 participants was not recorded) drawn from the same population pool as Experiment I were tested. All were native English speakers aged between 17-31 years with no history of hearing or language impairment. Seventy-five participants were right-handed and five participants were left-handed (handedness data for ten participants was not available). All participants had normal or corrected-to-normal vision and had not been previously exposed to NV speech. Subjects received either course credit or monetary compensation for their participation. The study was cleared by the Queen's University General Research Ethics Board.

Design and Materials

Test sentences were identical to Experiment I. I recorded an additional four sets of five sentences (for a total of 20 extra sentences; Appendix C) for use as training sentences using the same procedure noted in the previous experiment. Sentences were vocoded and processed as in Experiment I. The video camera audio file was extracted from the audiovisual file using Virtual Dub and was used as a template to align the normalized NV audio file in time (rather than the normalized clear audio file, as done in Experiment I). The degraded audio file was then recombined with the video file using Virtual Dub in order to produce the degraded audiovisual stimuli for the D_v and D_vD conditions.

Procedure

The 20 sentences of the training phase were presented in a fixed order to all participants; however, the format of presentation differed across conditions. Two groups were presented with degraded sentences with corresponding facial movements (D_v); one of these groups also received a second degraded (auditory-alone) presentation (D_vD). Participants in the CD group received an initial clear auditory-alone presentation of the sentence followed by the degraded form. Participants in the CDD group received the initial clear auditory-alone presentation of the sentence, followed by two repetitions of the degraded form. In all conditions, report for training sentences was collected following the first degraded presentation (Figure 6), although these are not really comparable across training conditions (for example, reporting the degraded sentence in the CD and CDD conditions is trivial, since subjects can simply report the C sentence).

After training, all four trained groups and a naïve (untrained) control group were tested using DC_vD trials similar to those in Experiment I (only the sentences used were

different). In this condition, an initial degraded auditory-only presentation is followed by the clear audiovisual version of the utterance, and then the degraded auditory form again. In all conditions, participants reported all the words they could understand after the first degraded presentation.

Following the 20 training sentences, the first twelve subjects in the DvD condition received the same counterbalanced 30 test sentences as subjects in Experiment I in DCvD format because I initially wanted to compare their performance with that of subjects in Experiment I. This format was altered for the last eight subjects in the DvD condition such that only two of the six sentence sets were used for each subject, for a total of ten sentences (rather than 30) presented in DCvD format. In order to combine the data from both groups of DvD participants, only the first ten sentences from the first group of twelve participants were included in analysis. All other subjects were tested on the same ten test sentences in DCvD format, with report following the first degraded sentence presentation (see Figure 6). The two sentence sets of the DCvD test were counterbalanced across listeners such that each sentence set was presented an equal number of times during the first and second set of five DCvD test sentences.

Prior to the start of the experiment, all subjects were told that the experiment had two phases. They were also informed of the format of the training and test trials in the experiment. There was a short delay following the training portions in order to signal to subjects that they were about to begin the second (test) phase. Instructions on the computer screen reminded subjects that they would hear each sentence in the test phase presented three times. Subjects were given 25 seconds for report in both the training and testing phases. In order to ascertain that working memory capacity did not limit subjects' performance on the subsequent tasks, subjects were prompted to type out four sentences

(6 to 10 words, M = 8 words/sentence, see Appendix B), presented as clear speech prior to the start of training, as in Experiment I. (All participants obtained perfect scores on the memory-test sentences, indicating that working memory capacity was not a limiting factor in their report scores). Subjects then heard a single sample sentence vocoded using fifteen bands, as an example of the form of distortion that they would hear in the task. A secondary speech-in-noise task was administered to all participants both prior to and following the experiment. Preliminary analysis indicates that the findings are unrelated to the results of this experiment. Thus, this task and the results will not be reported here.

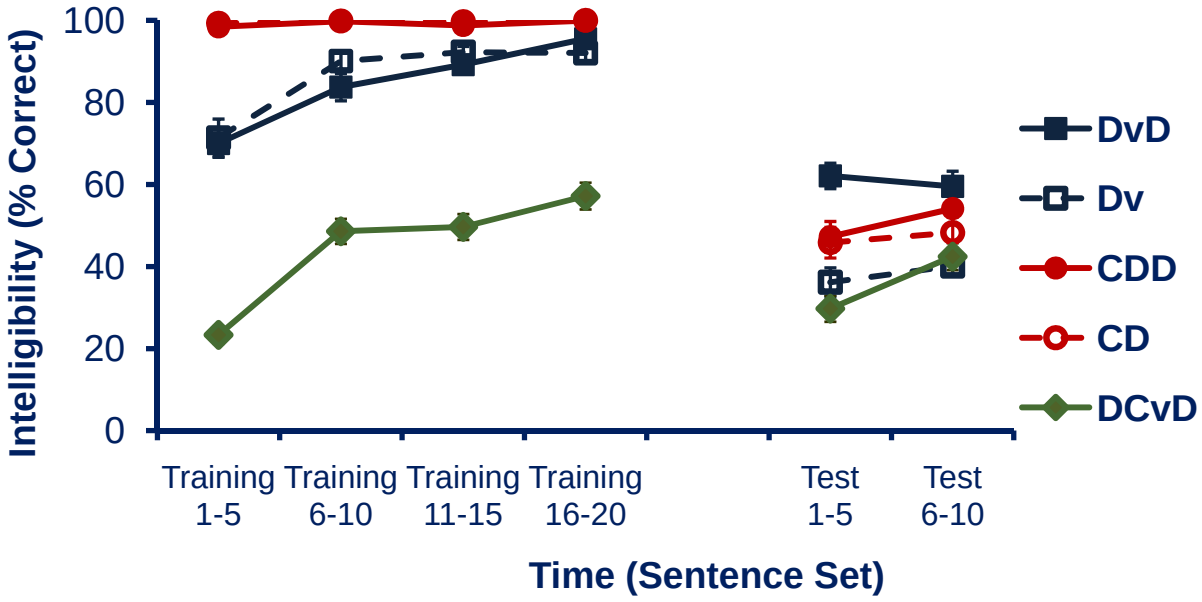
Data Coding and Analysis

Results were scored the same way as in Experiment I. Test scores were averaged over participants and items, identical to Experiment I, in order to confirm that any change in report scores over time and between conditions did not result from differences in sentence difficulty. I conducted 2-factor ANOVAs on word-report scores for test sentences presented in DCvD format, with Condition (5 levels) and Time (2 levels), averaging over participants (for the items analysis) and over items (for the analysis by participants). An additional dummy variable in the items analysis codes for which group of participants was tested first on each group of sentences, although main effects and interactions involving this dummy variable will not be reported (Pollatsek & Well, 1995). Condition (DCvD, Dv, DvD, CD, CDD) was entered as a between-subjects factor in the analysis by participants, and as a within-subjects factor in the analysis by items.

Results

There was a main effect of time, indicating improved report scores between the first five test trials of DCvD and the second five test trails. $F_1(1, 85) = 7.726$, $p = .007$, $\eta_p^2 = .083$; $F_2(1,24) = 10.697$, $p = .003$, $\eta_p^2 = .308$ (see Figure 7). There was also a main effect of condition, by both items and participants $F_1(4,85) = 13.822$, $MS = .362$, $p < .001$, $\eta_p^2 = .394$; $F_2(4,21) = 14.774$, $p < .001$, $\eta_p^2 = .738$, indicating that the presence and type of training had a significant effect on subjects' performance (Figure 7).

Figure 7. Experiment II average report scores for training and test sentences. Training and test scores performance for Experiment II, averaged over four groups of five training sentences and two groups of five test sentences in five groups of participants. Error bars show plus or minus one standard error of the mean over participants. Note that the overlap between scores for participants in the CD and CDD conditions during the training phase.



To analyze the effect of condition further, I performed post-hoc analysis using a Sidak correction for multiple comparisons. Analysis showed that performance in the DvD condition (M = 60.8%, SD = 14.5%, range = 33.9- 76.4%) was significantly greater than in the Dv condition (M = 38.1%, SD = 11.6%, range = 22.5- 56.9%, $p_1 < .001$; $p_2 < .001$). This indicates that repetition of the degraded sentence following an initial degraded presentation coupled with facial gestural information yields better performance than a single audiovisual degraded presentation. Performance in the DvD condition was also significantly greater than in the CD condition (M = 47.6%, SD = 14.1%, range = 33.7- 76.1%, $p_1 = .008$; $p_2 < .001$). Consistent with our hypothesis, subjects in the DvD condition also outperformed subjects in the baseline DCvD condition (M = 36.1%, SD = 13.7%, range = 16.4- 61.8%, $p_1 < .001$; $p_2 < .001$), indicating better learning from concurrent rather than sequential audiovisual information. Performance in the DvD condition was significantly better than in the CDD condition, by items (M = 50.1%, SD = 16.1%, range =32.8- 74.2%, $p = .016$), with a trend in the by-participants analysis (p = .090). This indicates that participants presented with concurrent facial gestural information outperformed subjects who received the clear version of the utterance followed by two degraded sentence presentations.

Performance did not differ significantly between subjects in the CDD and CD conditions, indicating that a second degraded presentation following presentation of the clear stimulus does not confer any additional advantage in learning, unlike the difference between DvD and Dv. Subjects in the CDD and CD conditions outperformed subjects in the DCvD condition (CDD: $p_1 = .002$; $p_2 = .006$; CD: $p_1 = .033$; $p_2 = .035$). Subjects in the CDD condition also outperformed those in the Dv condition ($p_1 = .014$; $p_2 = .001$), whereas for the CD condition, this difference was marginally not significant only by

items (p = .051). Performance for subjects in the D_v condition was not significantly different from subjects in the DC_vD condition. The Condition x Time interaction was not significant by participants or by items.

With respect to the 20 training sentences, subjects in the CDD and CD conditions reported nearly all words correctly (CDD: M = 99.22%, SD = 0.74%, range = 96.8-100%; CD: M = 99.70%, SD = 0.32%, range = 98.9-100%), indicating that subjects were paying attention. As evident in Figure 7, training report scores for the D_v and D_vD conditions, averaged across participants and items, were nearly identical and higher than any scores previously reported for NV speech (D_vD: M = 84.62%, SD = 8.20%, range = 67.8-96.4%; D_v: M = 86.44%, SD = 7.54%, range = 72.4- 96.5%). Despite differences in performance on test sentences between for subjects in the D_v and D_vD conditions, the two groups performed comparably on training sentences. Subjects in the DC_vD condition reported 44.71% of words from training sentences correctly (SD = 8.17%).

Discussion

In Experiment II, I demonstrated perceptual learning for degraded sentences presented with concurrent audiovisual feedback. I also demonstrated that this learning was enhanced when a second degraded sentence was provided. Subjects in the D_vD condition who received simultaneous audiovisual feedback for degraded sentence presentations outperformed subjects in the DC_vD who received clear audiovisual feedback after hearing degraded sentences. Thus, facial gestural information appears to be most useful to listeners when it is presented concurrently with the degraded acoustic utterance, especially when this is paired with a second degraded sentence repetition.

Interestingly, subjects in the D_v and D_vD conditions performed very similarly on training sentences, despite differences in performance on the test sentences. Degraded training sentences presented with concurrent facial information were highly intelligible to listeners; in both conditions, starting performance for training sentences was around 70% and often quickly reached 80-90% within the first ten sentences. Yet subjects in the D_vD condition dramatically outperformed those in the D_v condition on degraded test sentences presented in the auditory modality alone (in DC_vD format). This indicates that the second degraded sentence in the D_vD presentation format is crucial for generalization of learning to novel degraded utterances. In other words, although subjects in the D_v condition were able to directly map the facial gestural information onto the highly intelligible degraded sentence, the results indicate that it is necessary to hear the degraded sentence presentation a second time in order to produce a high degree of generalization to new sentences. This finding has important implications for rehabilitative programs for cochlear implant users and will be elaborated upon in the general discussion.

The results of the two control conditions offer insight into the factors that drive superior performance in the D_vD condition. The CDD condition also included two presentations of the degraded sentence (D), just like D_vD – however, unlike D_vD, the CDD condition provides subjects with complete linguistic information prior to hearing both degraded sentence presentations. Training sentences presented in this format are also readily intelligible and thus allow us to examine the influence of pop-out on perceptual learning; if pop-out truly drives learning, then performance in the CDD condition (where both degraded sentence presentations are completely intelligible) should have been greater than performance in the D_vD condition. Thus, these findings suggest that perceptual pop-out is not the only factor responsible for perceptual learning. Also,

since performance in the DvD group was significantly greater than performance in the CDD group in the analysis by items, with a trend towards significance in the analysis by participants, I may rule out the possibility that subjects' superior performance in the DvD, compared to Dv condition is due to exposure to a greater number of NV speech presentations. Instead, it appears that the presentation of concurrent facial gestural information coupled with repetition of the degraded sentence results in more efficient tuning of the perceptual apparatus.

The concurrent presentation of the degraded sentence with facial gestural information also did not account for the superior learning in the DvD condition when tested on auditory-alone degraded sentences. If presenting the degraded sentence concurrently with facial gestural information is solely responsible for optimal perceptual learning, then subjects in the Dv condition should outperform subjects in the CD condition. However, subjects in the CD condition marginally significantly outperformed subjects in the Dv condition by items (and this difference was not significant by participants). This indicates that the benefit of presenting concurrent facial gestural information for degraded sentences requires that the degraded speech form be presented a second time. Thus, it seems that learning is best facilitated when listeners map what they can render intelligible of the degraded audiovisual presentation onto a second degraded presentation when tested on auditory-alone degraded sentences.

This apparent importance of the second degraded sentence presentation implies that repetition is important for learning. This initially appears to be at odds with the lack of significant difference in performance between the CD and CDD groups; the data indicate that the second degraded presentation in the CDD condition does not confer any additional learning benefit. Perhaps a second degraded presentation is useful only if the

listener is not provided with complete linguistic information for the first degraded presentation, since they would be more likely to gain from the degraded sentence repetition.

Thus, the results suggest that superior performance in the DvD condition is not solely due the presence of the facial gestural information that accompanies the degraded sentence. However, an additional degraded presentation by itself does not contribute to this benefit in learning since improvements in performance are observed only under conditions in which incomplete linguistic information is supplied. It is a combination, or interaction, of these factors that produces the superior performance noted in the DvD condition: a combination of listeners' effortful processing in using facial gestures to help disambiguate the degraded sentence, and mapping hypotheses about the lexical content of the degraded sentence onto the degraded repetition.

In order to explore these interpretations in greater detail, I performed a 2-factor ANOVA for test sentences in the Dv, DvD, CD and CDD conditions. The linguistic information (complete/auditory, as in the CD and CDD conditions or incomplete/facial gestural, as in the Dv and DvD conditions) and number of degraded presentations (one or two) were both entered as between-subjects factors, with performance at each time point as the dependent measure. There was a main effect of the number of degraded presentations, $F(1, 68) = 22.357$, $p < .001$, $\eta_p^2 = .247$, indicating that repetition of degraded sentences aids performance. There was no main effect of linguistic information. As expected, the interaction between linguistic information and number of degraded presentations was significant, $F(1, 68) = 12.918$, $p = .001$, $\eta_p^2 = .160$, such that the second degraded presentation is most useful when concurrent facial gestural information (or incomplete linguistic information) is provided for the initial degraded presentation. The

results also confirm our interpretation of the main analysis; presenting facial gestural information by itself does not appear to enhance perceptual learning of degraded speech. These results will be discussed further in the general discussion.

It is also interesting that subjects in the CD condition outperformed subjects in the DCvD condition, despite the extra degraded sentence presentation in the latter condition. Why would the additional degraded presentation prior to receiving clear audiovisual feedback in the DCvD group effectively hinder performance? A follow-up experiment is necessary to explore these results, although I speculate that subjects' initial, erroneous, hypotheses about the identity of the degraded sentence in the DCvD condition may have interfered with their ability to map the clear speech form onto the second degraded sentence presentation, impeding their performance.

CHAPTER 4

GENERAL DISCUSSION

Why is Learning Best in DvD?

The aim of these two experiments was to evaluate the contribution of facial gestural information to perceptual learning of NV speech. Both Experiments I and II replicated two general findings noted in other studies of perceptual learning of NV speech: Listeners` performance improves over time and is further enhanced through feedback. In Experiment I, I demonstrated that presenting audiovisual clear feedback for degraded sentences does not confer any additional learning benefit over presenting auditory-alone clear feedback. In Experiment II, I demonstrated that facial gestural information results in the best perceptual learning of NV speech when this information is presented concurrently with degraded sentences, but only when this is followed by a second (auditory only) presentation of the degraded sentence. Taken together, the results substantiate a contribution for facial gestural information to perceptual learning of NV speech; however, this benefit occurs only under circumstances that allow for synchronous processing of visual and degraded auditory information *and* repetition, which leads to more efficient re-tuning of acoustic-phonetic links.

The interaction between the linguistic information provided and the number of degraded sentence presentations suggests that both factors may account for superior learning in the DvD condition, but only in combination. Indeed, if facial gestural information was essential for superior learning, then I should have seen comparable test performance for subjects in the DvD and Dv conditions since concurrent visual information for degraded sentences was presented to both groups. Instead, the presence of

facial gestural information may play a supporting role in producing the greater learning observed for the D_vD condition, possibly through recruitment of motor templates, helping to further constrain perceptual hypotheses (Callan, et al., 2003; Skipper, et al., 2005; Skipper, van Wassenhove, Nusbaum, & Small, 2007).

Davis et al. (2005) and Hervais-Adelman et al. (2008) proposed that better perceptual learning may be attributable to a process by which the degraded speech is mapped onto the clear acoustic form in order to tune perceptual representations. Such a proposal implies that the conditions that are thought to facilitate pop-out are also instrumental for learning. However, the results of our experiment challenge this interpretation; CDD, a condition for which complete pop-out should be observed, was less effective than D_vD, a condition which does not allow for complete pop-out, since the D_v speech is not 100% intelligible (Figure 7). Thus, better perceptual learning does not correspond to the extent and quality of feedback received, nor does it depend solely on the intelligibility of degraded sentences after feedback.

Although the degraded-speech components of the CDD (and CD) conditions are ~100% intelligible, the CDD condition was less conducive to learning than was the D_vD condition. This indicates that maximal levels of intelligibility do not necessarily produce better learning of degraded speech. However, levels of intelligibility of degraded speech that are too low also result in relatively poor learning, as, for example, in the DDC condition in Experiment I. Thus, there appears to be some 'sweet spot' at intermediate levels of intelligibility that appears to maximize perceptual learning of degraded speech.

The idea of optimal intelligibility for learning is consistent with evidence suggesting that the multisensory-integration system for speech is maximally tuned for SNRs where auditory and visual information are most likely to be supplementary to one

another. Gains in multisensory integration from viewing facial gestural information have been shown to be maximal at intermediate SNRs where recognition accuracy in the auditory-alone modality was approximately 20% (Ma, Zhou, Ross, Foxe, & Parra, 2009; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). Interestingly, the recognition accuracy levels reported by Ross et al. are similar to the initial levels of baseline word report accuracy noted in our study. The average starting performance for the first five sentences in the DDC, DCD and DC$_v$D conditions (conditions in which the initial degraded sentence for report was presented in the auditory modality alone) ranged from 18-26%. If I take into account that some learning occurs within the first five sentences, these levels approach the 20% recognition levels outlined by Ross et al. Although I use NV speech, whereas Ross et al. use speech in noise, the principle suggests that subjects trained in the D$_v$ and D$_v$D conditions, in which facial gestural information is supplied, may be optimally positioned for maximum gains in audiovisual integration, a topic that warrants further investigation.

Since listeners are performing best in a condition in which incomplete linguistic information is provided and I have ruled out any primary contributions of facial gestural information, this suggests that learning benefits observed in the D$_v$D condition may also be related to effort, or the degree to which listeners must work to extract linguistic information from the stimulus presentation. Whereas the CDD control condition provides listeners with complete linguistic information, the D$_v$D condition demands greater cognitive effort from listeners in order to extract the linguistic information from the combined facial gestures and degraded utterance. Educational studies have shown benefits for active versus passive learning (see Prince, 2004 for a review) and such an account is also consistent with documented contributions of attention to perceptual

learning (Ahissar & Hochstein, 1993; Fernandes, Kolinsky, & Ventura, 2010; Toro, Sinnett, & Soto-Faraco, 2005). Thus, more effortful engagement with the stimulus could account for the pattern of results, perhaps allowing listeners to further augment their learning (Seitz & Dinse, 2007).

Taken together, the data suggest that observed gains in perceptual learning are greatest under an optimal level of intelligibility and effort. In evaluating the plausibility of both of these explanations, however, I must also consider the importance of the second degraded sentence presentation when listeners have incomplete linguistic information. Similar levels of performance during training were noted for the D_v and D_vD conditions, but these two training conditions yielded dramatically different test performance. One possibility is that both groups may have quickly reached asymptotic performance levels in the training phase, and differences in learning would not emerge until subjects were presented with less intelligible (degraded, auditory-only) stimuli.

A repetition of the degraded sentence may promote learning for one of two reasons. First, whereas the first degraded presentation allows listeners to deduce the linguistic intent of the utterance, the mapping process and retuning of acoustic-phonetic links may occur upon the second degraded presentation. Second, it is possible that although some retuning occurs during the first degraded presentation, the subsequent degraded presentation allows listeners to further retune lower-level processes. Listeners may employ top-down strategies to utilize what they could understand from the first degraded presentation to decode more words in the second presentation. This might facilitate the development of richer and more complete perceptual hypotheses, simultaneously allowing listeners to test their initial perceptual hypotheses.

Alternatively, it is possible that the pattern of results observed between the training and testing phases for the two conditions is not due to differences in learning between the groups, but rather, the degree to which learning generalizes to a novel presentation format. In the DCvD test format, the initial degraded sentence is audio-only, without any accompanying visual information. The second degraded presentation allowed subjects in the DvD condition to obtain experience with a degraded audio-only stimulus, whereas subjects in the Dv condition did not have this opportunity. In a functional connectivity study, Nath and Beauchamp (2011) demonstrated differential weighting of information from different modalities depending on the fidelity of the signal, such the more reliable modality was more heavily weighted (see also Alais & Burr, 2004; Ernst & Banks, 2002; Ma, et al., 2009). Perhaps when a degraded sentence is presented concurrently with facial gestural information, as in the Dv and DvD conditions of our study, some weight is allocated to the facial gestural information at the expense of the auditory modality. The second degraded presentation in the DvD condition may allow for an adjustment or re-shifting of weight towards the auditory modality, resulting in better tuning of low-level auditory speech perception mechanisms. Thus, subjects in the DvD condition may be better equipped to comprehend the auditory-only degraded presentation of test sentences than subjects in the Dv condition, suggesting that the modality to which speech information is presented plays an important role. Further research is needed to explore this hypothesis, possibly by measuring report after both degraded presentations in the DvD condition, or by testing a DvDv condition (in which facial gestural information accompanies both degraded sentence presentations).

Other Findings

In line with the literature discussed in the Introduction, our results provide some support for the assertion that speech is represented amodally (e.g., Bernstein, et al., 2004; Liberman & Mattingly, 1985; Summerfield, 1987). Since I did not see any learning benefit for presenting audiovisual feedback rather than auditory-only feedback, I can rule out the possibility that speech is mapped onto visual articulatory representations, although it remains possible that speech representations might be auditory in nature.

Performance in the DCvD condition differed somewhat between Experiments I and II. In particular, the drop in performance for DCvD subjects between the training and test sentences in Experiment II was not observed in Experiment I. This difference is likely attributable to material effects. The 20 training sentences used in Experiment II were not used in Experiment I and may be easier, as evidenced by report scores of 25-50% for the first 10 sentences in Experiment II, compared to scores of 25-35% for the first 10 sentences presented in the same format in Experiment I. However, this drop in performance in Experiment II does not affect the interpretation of the main results.

Generalization

One limitation of our study is that I used a single talker and thus did not examine generalization to novel talkers. Generalization to a novel talker has been observed for in-depth training on one talker and also when listeners are trained on multiple talkers (Bradlow & Bent, 2008; Kraljic & Samuel, 2006), although it seems that the driving factor behind generalization is the degree to which the training stimuli capture the underlying variability present in the signal (Greenspan et al., 1988). Unpublished data from our group demonstrate that perceptual learning of NV speech generalizes to a novel

talker of the same accent; those trained on one talker and tested with another talker perform equally well as those who had been trained on that same talker. The data also show that additional experience with a novel talker can lead to further improvements, beyond those attained by those who were trained on only one talker for all trials. This is a crucial point for cochlear implant rehabilitation and warrants further investigation; exploration of this question will further help to optimize rehabilitative strategies for cochlear implant users.

An additional limitation is that in Experiment II, I tested generalization of participants' knowledge of NV speech using a DCvD feedback format, since I initially wanted the results to be comparable with Experiment I. A simpler and more direct test of generalization would be to test participants on auditory-only degraded sentences.

Individual Variability

Importantly, I noted a very high degree of individual variability with respect to the rate and degree of perceptual learning in both of our experiments, similar to Davis et al (2005). Individuals vary greatly in their ability to speechread (Dodd, Plant, & Gregory, 1989; Heider & Heider, 1940; MacLeod & Summerfield, 1987) and the same variability may apply to the extent to which individuals can extract facial gestural information for integration with the acoustic signal for lexical identification. Within the speechreading literature, there has been a concerted effort to uncover the root of superior speechreading abilities. Many cognitive factors have been implicated, including the ability to generate complete representations from partial information, verbal inference-making ability, working memory (i.e., the ability actively monitor and manipulate information needed to do complex tasks), processing speed (i.e., the speed at which an individual can process

information), low-level visual abilities and to a limited extent, intelligence and verbal abilities, although many of these findings remain controversial (see Summerfield, 1992 for a review). Future research should explore contributions to individual variability in learning of NV speech and cochlear implant rehabilitation, and whether these abilities can be improved through training. This line of research is particularly relevant for predicting outcomes for cochlear implant users in older adults (Sommers, Tye-Murray, & Spehar, 2005; Tye-Murray, Sommers, & Spehar, 2007).

Applications to Rehabilitation of Cochlear-Implant Users

Several practical strategies for rehabilitation of individuals with hearing loss can be extrapolated from my findings, particularly for the growing population of post-lingually deafened adults. I have studied the conditions that optimize perceptual learning for degraded speech, and their translational application may alleviate the psychosocial burdens associated with age-related hearing loss.

The translational potential of this work may be limited by differences between speech as transduced by a cochlear implant and NV speech. The signal in cochlear implants is a train of amplitude modulated electrical pulses, whereas NV speech uses a noise carrier (not a pulse train) and is acoustic. In addition, the snail-shaped anatomy of the cochlea often prevents complete insertion of the cochlear implant electrode array into the cochlea, resulting in a basalward spectral shift and a place-frequency mismatch. Finally, older cochlear implant wearers may have been deaf for many years, and their auditory systems may be physiologically abnormal, unlike those of the normal listeners tested here. All of these factors may account for the longer time-scale of learning for cochlear implant users, compared to the normal hearing participants used in this study.

Despite these drawbacks, studies of NV speech are an established method for simulating the conditions of listening to speech with an implant (Faulkner, Rosen, & Smith, 2000; Loizou, Dorman, & Tu, 1999; Shannon, et al., 1995),

The results of our experiments demonstrate that conditions allowing for extraction of information from both auditory and visual modalities may facilitate learning of novel-sounding speech more than the conditions employed by Davis et al (2005). Rehabilitative programs should focus on allowing listeners to make use of bimodal cues in speech perception; cochlear implant users should be encouraged to look at the speaker's face in order to exploit the rich source of articulatory speech cues available to them. My results also suggest that learning is optimized when listeners are provided with a second degraded presentation of each utterance. Thus, rehabilitative efforts could focus on devising software applications that enable listeners to hear sentence repetitions. In line with Ross et al. (2007), it might be efficacious to adjust the SNR to aim for 20% accuracy for auditory-alone presentations during training, either by reducing the level or adding noise, in order to allow cochlear-implant users to derive maximum benefit from bimodal speech cues. All of these suggestions represent potentially fruitful questions for future research.

Conclusion

The findings of these two experiments outline a limited role for facial gestural information in facilitating perceptual learning of degraded speech. I observed the greatest amount of learning in a condition that allowed for bimodal extraction of speech information. However, the importance of the second degraded presentation in supporting learning under these conditions indicates that the learning benefits with concurrent facial

gestural information cannot solely be explained by a privileged status for facial gestural information in the speech perception system. Instead, the additional benefit conferred by facial gestural information may be related to the greater effort that listeners must exert to extract linguistic information from the audiovisual degraded presentation. The pattern of results may also reflect an optimal level of intelligibility that is produced when facial gestural information is combined with the degraded speech form, when gains obtained from audiovisual integration are maximized. Further work is needed to explore the role of intelligibility and effort in perceptual learning of degraded speech, as these factors have important implications for the rehabilitation of cochlear implant users.

REFERENCES

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proc Natl Acad Sci U S A, 90*(12), 5718-5722.

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends Cogn Sci, 8*(10), 457-464.

Alais, D., & Burr, D. (2004). No direction-specific bimodal facilitation for audiovisual motion detection. *Brain Res Cogn Brain Res, 19*(2), 185-194.

Allen, N. H., Burns, A., Newton, V., Hickson, F., Ramsden, R., Rogers, J., . . . Morris, J. (2003). The effects of improving hearing in dementia. *Age Ageing, 32*(2), 189-193.

Altmann, G., & Young, D. (1993). Factors affecting adaptation to time-compressed speech. *EUROSPEECH'93*, 333-336.

Arlinger, S. (2003). Negative consequences of uncorrected hearing loss - a review. *International Journal of Audiology, 42*, S17-S20.

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology, 92*, 339-355.

Baart, M., & Vroomen, J. (2010). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neurosci Lett, 471*(2), 100-103.

Bent, T., Loebach, J. L., Phillips, L., & Pisoni, D. B. (2011). Perceptual adaptation to sinewave-vocoded speech across languages. *J Exp Psychol Hum Percept Perform*.

Bernstein, L. E., Auer, E. T., Jr., & Moore, J. K. (2004). Audiovisual Speech Binding: Convergence or Association? In G. A. Calvert, C. Spence & B. E. Stein (Eds.),

The handbook of multisensory processes (pp. 203-223). Cambridge, MA, US: MIT Press.

Bernstein, L. E., Auer, E. T., Moore, J. K., Ponton, C. W., Don, M., & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *Neuroreport, 13*(3), 311-315.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur J Neurosci, 20*(8), 2225-2234.

Boothroyd, A. (2007). Adult aural rehabilitation: what is it and does it work? *Trends Amplif, 11*(2), 63-71.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729.

Callan, D. E., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2001). Multimodal contribution to speech perception revealed by independent component analysis: a single-sweep EEG case study. *Brain Res Cogn Brain Res, 10*(3), 349-353.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport, 14*(17), 2213-2218.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., . . . David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*(5312), 593-596.

Carabellese, C., I, A., Rozzini, R., Bianchetti, A., Frisoni, G. B., Frattola, L., & Trabucchi, M. (1993). Sensory Impairment and Quality-of-Life in a Community Elderly Population. *Journal of the American Geriatrics Society, 41*(4), 401-407.

Carlsson, P. I., Hall, M., Lind, K. J., & Danermark, B. (2011). Quality of life, psychosocial consequences, and audiological rehabilitation after sudden sensorineural hearing loss. *Int J Audiol, 50*(2), 139-144.

Ching, T. Y. C., Dillon, H., & Byrne, D. (1998). Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification. *Journal of the Acoustical Society of America, 103*(2), 1128-1140.

Clark, G. M. (2003). Rehabilitation and Habilitation. In R. T. Beyer (Ed.), *Cochlear implants: Fundamentals and applications* (pp. 654-706). New York: Springer-Verlag.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *J Acoust Soc Am, 116*(6), 3647-3658.

Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Lang Speech, 47*(Pt 3), 207-239.

Cohen, S. M., Labadie, R. F., & Haynes, D. S. (2005). Primary care approach to hearing loss: the hidden disability. *Ear Nose Throat J, 84*(1), 26, 29-31, 44.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clin Neurophysiol, 113*(4), 495-506.

D'Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2011). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex*.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hear Res, 229*(1-2), 132-147.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *J Exp Psychol Gen, 134*(2), 222-241.

Dodd, B., Plant, G., & Gregory, M. (1989). Teaching lip-reading: the efficacy of lessons on video. *Br J Audiol, 23*(3), 229-238.

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *J Exp Psychol Hum Percept Perform, 23*(3), 914-927.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Percept Psychophys, 67*(2), 224-238.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429-433.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur J Neurosci, 15*(2), 399-402.

Fahle, M. (2005). Perceptual learning: specificity versus generalization. *Curr Opin Neurobiol, 15*(2), 154-160.

Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: implications for cochlear implants. *J Acoust Soc Am, 108*(4), 1877-1887.

Fernandes, T., Kolinsky, R., & Ventura, P. (2010). The impact of attention load on the use of statistical information and coarticulation as speech segmentation cues. *Atten Percept Psychophys, 72*(6), 1522-1532.

Fine, I., & Jacobs, R. A. (2002). Comparing perceptual learning tasks: a review. *J Vis, 2*(2), 190-203.

Fishman, K. E., Shannon, R. V., & Slattery, W. H. (1997). Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor. *J Speech Lang Hear Res, 40*(5), 1201-1215.

Fitzgerald, M. B., & Wright, B. A. (2005). A perceptual learning investigation of the pitch elicited by amplitude-modulated noise. *J Acoust Soc Am, 118*(6), 3794-3803.

Francis, H. W., Chee, N., Yeagle, J., Cheng, A., & Niparko, J. K. (2002). Impact of cochlear implants on the functional health status of older adults. *Laryngoscope, 112*(8), 1482-1488.

Fu, Q. J., & Galvin, J. J., 3rd. (2003). The effects of short-term training for spectrally mismatched noise-band speech. *J Acoust Soc Am, 113*(2), 1065-1072.

Fu, Q. J., & Galvin, J. J., 3rd. (2007). Perceptual learning and auditory training in cochlear implant recipients. *Trends Amplif, 11*(3), 193-205.

Godde, B., Stauffenberg, B., Spengler, F., & Dinse, H. R. (2000). Tactile coactivation-induced changes in spatial discrimination performance. *J Neurosci, 20*(4), 1597-1604.

Goldstone, R. L. (1998). Perceptual learning. *Annu Rev Psychol, 49*, 585-612.

Granger, R., & Lynch, G. (1991). Higher olfactory processes: perceptual learning and memory. *Curr Opin Neurobiol, 1*(2), 209-214.

Grant, K. W., & Braida, L. D. (1991). Evaluating the articulation index for auditory-visual input. *J Acoust Soc Am, 89*(6), 2952-2960.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J Acoust Soc Am, 108*(3 Pt 1), 1197-1208.

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.), Hearing by eye II: The psychology of speechreading and audiovisual speech. Hove, UK: Psychology Press.

Green, K. P., & Gerdeman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: the McGurk effect with mismatched vowels. *J Exp Psychol Hum Percept Perform, 21*(6), 1409-1426.

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. *Percept Psychophys, 50*(6), 524-536.

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Percept Psychophys, 38*(3), 269-276.

Green, K. P., & Norrix, L. W. (2001). Perception of /r/ and /l/ in a stop cluster: evidence of cross-modal context effects. *J Exp Psychol Hum Percept Perform, 27*(1), 166-177.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *J Exp Psychol Learn Mem Cogn, 14*(3), 421-433.

Greenwood, D. D. (1990). A Cochlear Frequency-Position Function for Several Species - 29 Years Later. *Journal of the Acoustical Society of America, 87*(6), 2592-2605.

Heider, F., & Heider, G. (1940). An experimental investigation of lipreading. *Psychological Monographs, 52*, 124-153.

Hervais-Adelman, A., Carlyon, R. P., Johnsrude, I. S., & Davis, M. H. (2011). *The dorsal speech pathway is recruited for effortful comprehension of noise-vocoded words: Evidence from fMRI. Manuscript submitted for publication*.

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: effects of feedback and lexicality. *J Exp Psychol Hum Percept Perform, 34*(2), 460-474.

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *J Exp Psychol Hum Percept Perform, 37*(1), 283-295.

Hochstein, S., & Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron, 36*(5), 791-804.

Horn, K. L., McMahon, N. B., McMahon, D. C., Lewis, J. S., Barker, M., & Gherini, S. (1991). Functional use of the Nucleus 22-channel cochlear implant in the elderly. *Laryngoscope, 101*(3), 284-288.

Iacoboni, M. (2008). The role of premotor cortex in speech perception: evidence from fMRI and rTMS. *J Physiol Paris, 102*(1-3), 31-34.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). "Putting the face to the voice": matching identity across modality. *Curr Biol, 13*(19), 1709-1714.

Kelsall, D. C., Shallop, J. K., & Burnelli, T. (1995). Cochlear implantation in the elderly. *Am J Otol, 16*(5), 609-615.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychon Bull Rev, 13*(2), 262-268.

Labadie, R. F., Carrasco, V. N., Gilmer, C. H., & Pillsbury, H. C., 3rd. (2000). Cochlear implant performance in senior citizens. *Otolaryngol Head Neck Surg, 123*(4), 419-424.

Lachs, L., & Pisoni, D. B. (2004). Cross-modal source information and spoken word recognition. *J Exp Psychol Hum Percept Perform, 30*(2), 378-396.

LaForge, R. G., Spector, W. D., & Sternberg, J. (1992). The relationship of vision and hearing impairment to one-year mortality and functional decline. *Journal of Aging and Health, 4*, 126-148.

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. *J Exp Psychol Hum Percept Perform, 33*(4), 905-914.

Li, L., Daneman, M., Qi, J. G., & Schneider, B. A. (2004). Does the information content of an irrelevant source differentially affect spoken word recognition in younger and older adults? *Journal of Experimental Psychology-Human Perception and Performance, 30*(6), 1077-1091.

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

Loebach, J. L., & Pisoni, D. B. (2008). Perceptual learning of spectrally degraded speech and environmental sounds. *J Acoust Soc Am, 123*(2), 1126-1139.

Loebach, J. L., Pisoni, D. B., & Svirsky, M. A. (2010). Effects of semantic context and feedback on perceptual learning of speech processed through an acoustic simulation of a cochlear implant. *J Exp Psychol Hum Percept Perform, 36*(1), 224-234.

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *J Acoust Soc Am, 106*(4 Pt 1), 2097-2103.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space. *PLoS One, 4*(3).

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *Br J Audiol, 21*(2), 131-141.

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24*(1), 29-43.

MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A. S., McGuire, P., . . . Brammer, M. J. (2000). Silent speechreading in the absence of scanner noise: an event-related fMRI study. *Neuroreport, 11*(8), 1729-1733.

MacSweeney, M., Calvert, G. A., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C., . . . Brammer, M. J. (2002). Speechreading circuits in people born deaf. *Neuropsychologia, 40*(7), 801-807.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: lexical adaptation to a novel accent. *Cogn Sci, 32*(3), 543-562.

McConkey-Robbins, A. (2000). Rehabilitation after cochlear implantation. In J. Niparko, K. Kirk, N. Mellon, A. McConkey, D. Tucci & B. Wilson (Eds.), *Cochlear*

implants: Principles and practices (pp. 323-367). Philadelphia: Lippincott, Williams & Wilkins.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

MedlinePlus. (2010). Age-related hearing loss Retrieved July 15, 2010, from http://www.nlm.nih.gov/medlineplus/ency/article/001045.htm

Mehler, J., Sebastian, N., Altmann, G., Dupoux, E., Christophe, A., & Pallier, C. (1993). Understanding Compressed Sentences - the Role of Rhythm and Meaning. *Temporal Information Processing in the Nervous System, 682*, 272-282.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr Biol, 17*(19), 1692-1696.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41*, 329-335.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: a reanalysis and some implications. *Phonetica, 41*(4), 215-225.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Percept Psychophys, 25*(6), 457-465.

Mulrow, C. D., Aguilar, C., Endicott, J. E., Velez, R., Tuley, M. R., Charlip, W. S., & Hill, J. A. (1990). Association between Hearing Impairment and the Quality of

Life of Elderly Individuals. *Journal of the American Geriatrics Society, 38*(1), 45-50.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychol Sci, 15*(2), 133-137.

Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *J Neurosci, 31*(5), 1704-1714.

NICD. (2009). Cochlear Implants Retrieved July 12, 2010, from http://www.nidcd.nih.gov/health/hearing/coch.asp

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cogn Psychol, 47*(2), 204-238.

Pallier, C., Sebastian-Galles, N., Dupoux, E., Christophe, A., & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: a cross-linguistic study. *Mem Cognit, 26*(4), 844-851.

Pasanisi, E., Bacciu, A., Vincenti, V., Guida, M., Barbot, A., Berghenti, M. T., & Bacciu, S. (2003). Speech recognition in elderly cochlear implant recipients. *Clinical Otolaryngology, 28*(2), 154-157.

Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I. P., Mottonen, R., Tarkiainen, A., & Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport, 16*(2), 125-128.

Pichora-Fuller, M. K., & Singh, G. (2006). Effects of age on auditory and cognitive processing: implications for hearing aid fitting and audiologic rehabilitation. *Trends Amplif, 10*(1), 29-59.

Poeppel, D., & Monahan, P. J. (2010). Feedforward and feedback in speech perception: Revisiting analysis by synthesis. *Language and Cognitive Processes, 1*, 1-17.

Pollatsek, A., & Well, A. D. (1995). On the Use of Counterbalanced Designs in Cognitive Research - a Suggestion for a Better and More Powerful Analysis. *Journal of Experimental Psychology-Learning Memory and Cognition, 21*(3), 785-794.

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93*(3), 223-231.

Pulvermuller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc Natl Acad Sci U S A, 103*(20), 7865-7870.

Recanzone, G. H., Merzenich, M. M., Jenkins, W. M., Grajski, K. A., & Dinse, H. R. (1992). Topographic reorganization of the hand representation in cortical area 3b owl monkeys trained in a frequency-discrimination task. *J Neurophysiol, 67*(5), 1031-1056.

Reisberg, D., McLean, J., & Goldfield, A. (Eds.). (1987). *Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Remez, R. E. (2005). Perceptual Organization of Speech. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception*. Oxford: Blackwell Publishing.

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.

Remez, R. E., Nygaard, L. C., Rubin, P. E., & Howell, W. A. (1987). Perceptual Normalization of Vowels Produced by Sinusoidal Voices. *Journal of Experimental Psychology-Human Perception and Performance, 13*(1), 40-61.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech-Perception without Traditional Speech Cues. *Science, 212*(4497), 947-950.

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cereb Cortex, 15*(8), 1261-1269.

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: implications for cochlear implants. *J Acoust Soc Am, 106*(6), 3629-3636.

Rosenblum, L. D. (1994). How Special Is Audiovisual Speech Integration. *Cahiers De Psychologie Cognitive-Current Psychology of Cognition, 13*(1), 110-116.

Rosenblum, L. D. (2008). Speech Perception as a Multimodal Phenomenon. *Current Directions in Psychological Science, 17*(6), 405-409.

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects. *Psychol Sci, 18*(5), 392-396.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environment. *Cerebral Cortex, 17*(5), 1147-1153.

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., & Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci Lett, 127*(1), 141-145.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Atten Percept Psychophys, 71*(6), 1207-1218.

Schneider, B. A., Daneman, M., & Pichora-Fuller, M. K. (2002). Listening in aging adults: From discourse comprehension to Psychoacoustics. *Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie Experimentale, 56*(3), 139-152.

Schwartz, J. L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech* (pp. 85-108). Hove, UK: Psychology Press.

Scott, D. R., & Cutler, A. (1984). Segmental Phonology and the Perception of Syntactic Structure. *Journal of Verbal Learning and Verbal Behavior, 23*(4), 450-466.

Sebastian-Gallés, N., Dupoux, E., Segui, J., & Mehler, J. (1992). Contrasting Syllabic Effects in Catalan and Spanish. *Journal of Memory and Language, 31*(1), 18-32.

Seitz, A. R., & Dinse, H. R. (2007). A common framework for perceptual learning. *Curr Opin Neurobiol, 17*(2), 148-153.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303-304.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *J Exp Psychol Hum Percept Perform, 28*(6), 1447-1469.

Shin, Y. J., Fraysse, B., Deguine, O., Vales, O., Laborde, M. L., Bouccara, D., . . . Uziel, A. (2000). Benefits of cochlear implantation in elderly patients. *Otolaryngology-Head and Neck Surgery, 122*(4), 602-606.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage, 25*(1), 76-89.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb Cortex, 17*(10), 2387-2399.

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *J Acoust Soc Am, 118*(5), 3177-3186.

Smith, D. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *J Acoust Soc Am, 117*(1), 305-318.

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear Hear, 26*(3), 263-275.

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America, 26*(2), 212-215.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *J Exp Psychol Hum Percept Perform, 7*(5), 1074-1095.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audiovisual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: the psychology of lipreading* (pp. 3-51). Hillsdale, NJ: Lawrence Erlbaum.

Summerfield, Q. (1992). Lipreading and Audiovisual Speech-Perception. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 335*(1273), 71-78.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97*(2), B25-34.

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear Hear, 28*(5), 656-668.

Uhlmann, R. F., Larson, E. B., Rees, T. S., Koepsell, T. D., & Duckert, L. G. (1989). Relationship of Hearing Impairment to Dementia and Cognitive Dysfunction in Older Adults. *Jama-Journal of the American Medical Association, 261*(13), 1916-1919.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci U S A, 102*(4), 1181-1186.

Watkins, K., & Paus, T. (2004). Modulation of motor excitability during speech perception: the role of Broca's area. *J Cogn Neurosci, 16*(6), 978-987.

Watkins, K., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*(8), 989-994.

Weill, S. A. (2001). Foreign accented speech: Adaptation and generalization. [Unpublished master's thesis]. *Ohio State University*.

WHO. (2001). Health and ageing: A discussion paper Retrieved July 17, 2010, from http://whqlibdoc.who.int/hq/2001/WHO_NMH_HPS_01.1.pdf

WHO. (2002). Active ageing: A policy framework. *A contribution of the World Health Organization to the Second United Nations World Assembly on Ageing* Retrieved July 17, 2010, from http://whqlibdoc.who.int/hq/2002/WHO_NMH_NPH_02.8.pdf

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat Neurosci, 7*(7), 701-702.

Appendix A

Test Sentences

The whole sky was full of birds.
The sketch showed that the road would pass the school.
The carpet and the curtains were the same colour.
The award was given to the writer at the end of his career.
The police returned to the museum.
Spiders are often found in the bath.
The fireman climbed down into the bottom of the tunnel.
He left school before he had done his exams.
It was a sunny day and the children were going to the park.
His new clothes were from France.
The coin was thrown onto the floor.
The burglar came up over the wall of the palace.
He reminded his parents about the game of football.
The group of friends got a taxi home after they left the nightclub.
His wig fell on the floor.
The man read the newspaper at lunchtime.
The camel was kept in a cage at the zoo.
The child left all of his lunch at home.
The children were hoping to play some hockey and rugby at their school.
The rice was cooked in a large saucepan.
The student tried to move the desk.
The soup was kept in a carton in the fridge.
The noise was very loud and difficult to ignore.
It was the crew that remained when the final lifeboat left the ship.
The car drove over the cliff.
Some ice was added to the whisky.
His face showed that his team had lost the game.
He was sitting at his desk in his office.
The furniture in the dining room was removed when the room was decorated.
There were books in the cellar.

Appendix B

Sentences Presented as Clear Speech

The author wrote the book that year.
The church was destroyed by the blaze.
The beef was rare, just as the customer had requested.
She arrived at the shop before it was open.

Appendix C

Training Sentences

Her new skirt was made of denim.
The gambler lost most of his money at the races.
He surprised his parents by his lack of concern.
It was the women that complained when the old bingo hall was closed.
The television program was a success.
The cattle were kept in the barn.
He broke his leg when he fell off the horse.
He always read a book before going to bed.
The dessert was put in the oven at the start of the meal.
The old tree was in danger.
The statue had some paint on it.
The goat was greedy just as the family had expected.
The knife was rather blunt and awkward to use.
The money for the science library was increased when the university was modernized.
There were mice in the cave.
Her secrets were written in her diary.
A spoon was used to stir the cup of tea.
The child was sad when her toys were damaged.
The woman was hoping to discover the name and address of the culprit.
The game ended as a draw.