The Thesis Committee for Han-Gyol Yi
Certifies that this is the approved version of the following thesis:

Audiovisual Integration for Perception of Speech Produced by Nonnative Speakers

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor: _____
Bharath Chandrasekaran

Supervisor: _____
Rajka Smiljanic

Audiovisual Integration for Perception of Speech Produced by Nonnative Speakers

by

Han-Gyol Yi, B.S.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

August 2013

Acknowledgements

This experiment was conducted in conjunction with Jasmine E. B. Phelps, Bharath Chandrasekaran, and Rajka Smiljanic. The research was funded by Longhorn Innovation Fund for Technology, awarded to BC and RS. Kristin J. Van Engen provided invaluable assistance in stimulus preparation and insight to data analysis. The following SoundBrain Lab research assistants collected the data: Blue Alozie, Tiffany Berry, Kadee Bludau, Millicent Campbell, Kathryn Curry, Dionne Dias, Morgan Elkins, Sarah Evans, Kim Kowinski, Carolyn Linebaugh, and Elsa Tran.

Abstract

Audiovisual Integration for Perception of Speech Produced by Nonnative Speakers

Han-Gyol Yi, M.A.

The University of Texas at Austin, 2013

Supervisors: Bharath Chandrasekaran and Rajka Smiljanic

Speech often occurs in challenging listening environments, such as masking noise. Visual cues have been found to enhance speech intelligibility in noise. Although the facilitatory role of audiovisual integration for perception of speech has been established in native speech, it is relatively unclear whether it also holds true for speech produced by nonnative speakers. Native listeners were presented with English sentences produced by native English and native Korean speakers. The sentences were in either audio-only or audiovisual conditions. Korean speakers were rated as more accented in audiovisual than in the audio-only condition. Visual cues enhanced speech intelligibility in noise for native English speech but less so for nonnative speech. Reduced intelligibility of audiovisual nonnative speech was associated with implicit Asian-Foreign association, suggesting that listener-related factors partially influence the efficiency of audiovisual integration for perception of speech produced by nonnative speakers.

Table of Contents

List of Figures

Chapter 1: Introduction

1. SUMMARY

In this thesis, I will examine the role of audiovisual integration in nonnative speech perception. In the Introduction section, I will review the current literature on audiovisual integration in speech perception and nonnative speech perception in noise. I will then present a study[1] that has been conducted recently examining this topic. The goal of the study was to compare the extent of beneficial effect of visual cues on perception of speech produced by native English and native Korean speakers. It was found that native listeners are less efficient in using nonnative visual cues to enhance speech intelligibility. Furthermore, the magnitude of the implicit association between East Asian faces and Foreignness predicted the enhanced native speech intelligibility when visual cues were available.

2. SPEECH PERCEPTION IN NOISE

Speech communication rarely takes place in an ideal setting. There are multiple factors that challenge speech processing. One of these is the impact of background noise. Extraneous auditory signals co-occurring with the speech signals can be detrimental to the target signal (Sumby & Pollack, 1954). The interference with the target signal can exist along a spectrum of two extremes. On one end of the spectrum is energetic masking, which competes with the speech signal at a peripheral level. Examples of this type of noise masker include the sound of passing cars, loud air vents, and construction noise. The other is informational masking, which masks the signal at a more central level. The

[1] Portions of the findings from this experiment have been published in Journal of the Acoustical Society of America – Express Letters (Yi, Phelps, Smiljanic, & Chandrasekaran, 2013).

presence of linguistic signal can confuse the listener via its semantic content unrelated to the target signal, in addition to the energetic masking arising from the acoustical energy of the informational masker (Lecumberri, Cooke, & Cutler, 2010; Mattys, Davis, Bradlow, & Scott, 2012). Examples of this type of noise masker involve one or more additional talkers producing speech as the target speech is being produced (Pollack, 1975). In the laboratory, energetic masking is studied by embedding the speech signal in a static wideband acoustic noise, while informational masking is studied by embedding the speech signal in a "babble" of multiple talkers. To minimize the effect of the noise on speech intelligibility, listeners use various strategies. A subset of these provide information regarding both the temporal onset of target speech production and the identity of the phoneme produced to enable effective stream segregation (Freyman, Balakrishnan, & Helfer, 2004; Kidd Jr, Mason, Deliwala, Woods, & Colburn, 1994; Kidd Jr, Mason, & Gallun, 2005; Parbery-Clark, Skoe, Lam, & Kraus, 2009). Therefore, speech perception in masking noise can benefit from cues that provide temporal or phonemic information (Grant & Seitz, 2000).

3. SPEECH AUDIOVISUAL INTEGRATION

Speech communication often occurs face-to-face. This means that visual cues that necessarily accompany speech production are also available to the listeners. These visual cues are known to benefit speech intelligibility in noise due to multiple reasons (Erber, 1975; Girin, Schwartz, & Feng, 2001; MacLeod & Summerfield, 1987, 1990; Sumby & Pollack, 1954). First, visual cues are immune to sources of acoustic noise. Second, visual cues inform the listener of precise temporal onset of speech sounds. This is especially beneficial for informational masking because it allows the listener to focus only on the speech sound that is synchronized with onset of visual speech production (Macaluso,

George, Dolan, Spence, & Driver, 2004; Summerfield, 1992; Vatakis & Spence, 2006). Third, visual cues provide general phonemic information of some speech sounds that are produced. Many consonants and vowels can be differentiated according to the shapes of speech articulators that are required to produce the individual sounds (Rosenblum & Saldaña, 1996; Schwartz, Berthommier, & Savariaux, 2004). Individuals vary in their ability to accurately perceive sentences presented only with visual cues, with hearing impaired population sometimes being able to reach more than 80% percent accuracy or as low as 11%, while young adults with no hearing problems not being able to reach more than 50% accuracy (Dodd, Plant, & Gregory, 1989; Heider & Heider, 1940; MacLeod & Summerfield, 1987; Summerfield, 1992). Regardless, normal hearing young adults can still use contextual information coupled with speech-reading to result in improved speech perception (Benguerel & Pichora-Fuller, 1982; Benoit, Mohamadi, & Kandel, 1994; Erber & McMahan, 1976; Matthews, Cootes, Bangham, Cox, & Harvey, 2002; Montgomery & Jackson, 1983; Montgomery, Walden, & Prosek, 1987; Rönnberg, Samuelsson, & Lyxell, 1998).

In understanding the role of visual cues in speech perception, three listening situations can be hypothesized. The first situation is an ideal listening environment without any external noise. Here, high levels of speech intelligibility can be attained even in the absence of visual cues, which implies that there is no additional benefit to be gained from the existence of visual cues. The second situation is a listening environment with an extreme degree of noise that completely masks the target speech signal. In such a situation, normal hearing young adults cannot retrieve the entirety of the speech stimuli with visual cues alone (Summerfield, 1992). In contrast with these two conditions, the third situation is a situation with moderate levels of acoustic noise that reduces speech intelligibility. Here, the auditory signal is degraded so that speech intelligibility is

compromised with auditory cues alone, but the level of degradation is sufficient to allow bootstrapping via visual cues to achieve significantly improved speech perception. This notion of the "sweet-spot" in which both the necessity and viability of visual cues in speech perception are maximized has been evidenced in several previous studies (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954).

A particularly well-studied audiovisual phenomenon is the so-called McGurk effect (McGurk & MacDonald, 1976). In their seminal paper, McGurk and MacDonald (1976) found that visual cues can modulate the percept of speech sounds. When the auditory bilabial stop consonant (e.g., /b/) is presented simultaneously with the visual velar stop consonant (e.g., /g/), listeners perceive the alveolar stop consonant (e.g., /d/), which does not exist in either the auditory or the visual streams (McGurk & MacDonald, 1976). Further research has revealed that the degree of visual interference on auditory perception of the stop consonants is modulated by the perceived relative reliability of either stream (Nath & Beauchamp, 2011). If noise is added to the auditory stream, the consonant is perceived with greater weighting on the visual stream. Conversely, noise added to the visual stream causes the perception to be biased towards the syllable in the auditory stream (computer screen covered with film; Fixmer & Hawkins, 1998; contrast and spatial resolution reduced; Nath & Beauchamp, 2011). Based on these observations, the authors conclude that audiovisual speech processing is dynamic, with weighting on the either modality malleable according to the perceived reliability. In other words, the goal of the listener is to perceive the incoming speech signal as accurately as possible. Consequently, perception is biased towards the source of the signal that is considered to be less degraded and more faithful to the original production (Massaro, 1998). According to this fuzzy logical model of perception (FLMP), inputs from the two modalities as the sole determinants of audiovisual perception of speech, while the integration process itself

is deemed to be universal and listener-invariant. However, McGurk susceptibility has been found to be affected by cultural differences, which questions the universality of this phenomenon (Sekiyama & Tohkura, 1993). Hence, the FLMP has been challenged recently, based on the findings that the patterns in audiovisual integration exhibit individual variability independent from patterns in unimodal perception (Schwartz, 2010). This subject-dependent audiovisual integration may not be exclusively described by simple weighting decisions on either modality, and additional factors may play into the end result in audiovisual speech perception.

4. The Challenge of Nonnative Speech

Noise is not the only source of degradation of the speech signal. Environmental noise compromises speech intelligibility from external sources. This means that the masking source can be isolated and its effects considered independently from the speaker. However, speaker-driven factors can also hinder speech comprehension. There are two main ways in which speaker intelligibility varies. First, speech intelligibility can vary within a speaker. Speakers modify speech styles depending on the situation. When speaking with familiar interlocutors or in a casual setting, speech tends to be in conversational style, which leads to fast speech with sound reductions and deletions (Picheny, Durlach, & Braida, 1985, 1986, 1989). When speaking with unfamiliar interlocutors or in a formal setting, speakers tend to speak in clear style, which leads to slower speech with more exaggerated sound enunciation (Bradlow & Bent, 2002; Ferguson, 2004; Ferguson & Kewley-Port, 2007; Helfer, 1997; Picheny, et al., 1985, 1986, 1989; Smiljanić & Bradlow, 2005). Additionally, speakers attempt to override the undesirable effects of environmental noise by modifying their speech to enhance communicative effectiveness (Junqua, Fincke, & Field, 1999; Lombard, 1911). Second,

speech intelligibility varies between speakers; some speakers produce more intelligible speech than do others (Hazan & Markham, 2004). In general, female speakers tend to be more intelligible than male speakers (Bradlow, Torretta, & Pisoni, 1996). Familiarity with the speakers increases speech intelligibility (Bradlow & Pisoni, 1999). Another source of between-talker variability in speech intelligibility is nonnative speech, in which the speaker is speaking in a language other than one's native language, or L1. Nonnative speakers of a language produce speech in a way that is perceptively deviates from the native targets. This difference often leads to the reduced intelligibility (Munro & Derwing, 1995a, 1995b; Rogers, DeMasi, & Krause, 2010; Rogers, Lister, Febo, Besing, & Abrams, 2006), especially in a more challenging listening situation, such as in noise (Munro, 1998).

As discussed earlier, visual cues can enhance speech intelligibility in environmental noise (Sumby & Pollack, 1954). In this regard, it could be conjectured that visual cues may play a similar role in nonnative speech perception, in which the degraded auditory signal leads to reduced intelligibility. Indeed, evidence suggests that native listeners of a language may place a greater weight on the visual stream of the speech signal when perceiving speech produced by nonnative speakers. In a recent study (Hazan, Kim, & Chen, 2010), native British and Australian English speakers listened to stop consonants (/ba/, /da/, /ga/) produced by native Australian English or Mandarin Chinese speakers. The stimuli contained conflicting auditory and visual consonant information. It was discovered that the native English listeners, regardless of their country of origin, were more likely to place greater weighting on the visual modality of the speech information for syllables produced by native Mandarin speakers than for those produced by native English speakers. This pattern did not exist in native Mandarin listeners. This finding suggests that the relative weighting of the auditory against visual speech

information is not only independent on the signal-driven degradation, but also influenced by the linguistic knowledge of the listeners. However, there still remain a few unresolved questions from these results.

The first issue concerns the generalizability of the perception of monosyllabic stimuli to that of words or sentences. Lexical contextual information is absent in syllable perception. It cannot be confidently supposed that the pattern of increased visual weighting on the nonnative speech stimuli will be replicated when there are additional semantic cues that could resolve ambiguity in the distorted signal (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005). Listeners, when these contextual cues are made available, may be less prone to having to rely on visual cues to enhance intelligibility.

The second issue concerns whether the increased visual weighting of nonnative speech stimuli is beneficial to accurate speech perception. In the Hazan et al. (2010) study, the participants were instructed to report the syllables that they "perceived". If the instruction had been to report the syllable "heard", then the increased visual weighting could be considered harmful to accurate perception in the case of incongruous audiovisual stimuli. On the contrary, if the instruction had been to report the syllable "seen", then the increased visual weighting could be considered beneficial to accurate perception. Since the instruction had not veered towards either of the unisensory modalities, the phenomenon of increased visual weighting is neutral in terms of being assessed of its effect on accurate perception. However, in a more realistic speech in noise perception situation, there exists a right answer. If increased visual weighting takes place in sentence comprehension despite degraded visual cues, there is no guarantee that it will actually enhance speech intelligibility. In fact, there is some evidence to suggest the contrary. First, it has been found that nonnative listeners are less efficient in using visual

cues to resolve ambiguity in the incoming speech signal (Hazan et al., 2006). If this is true, then native listeners less experienced in speech produced by nonnative speakers may be less efficient in using nonnative visual cues to enhance speech intelligibility. Although nonnative visual cues might be degraded relative to native visual cues, conflicting audiovisual cues such as in McGurk experiments do not occur in real-life speech. It would be reasonable to posit that although degraded visual cues may not be as beneficial in resolving degraded auditory cues, they would not negatively affect speech perception. Second, the beneficial effect of visual cues is modulated by the magnitude of enhancement of speech gestures. For instance, speakers exaggerate the facial motions required in speech production in noise. It has been found that audiovisual integration for speech produced in noise is more effective than for speech produced in quiet (Kim, Sironic, & Davis, 2011). This finding implies that the visual speech cues with less exaggeration of facial motions would induce less effective audiovisual integration, and that the conduciveness of the visual cues contributes to the overall effectiveness of audiovisual integration.

The third issue concerns the possible source of increased visual weighting. It has been found that listener-related factors affect the relative amount of visual weighting on the nonnative speech stimuli. When the audiovisual speech signal carries conflicting auditory vs. visual information, native listeners tend to rely more on the visual cues for nonnative speakers than for native speakers (Hazan, et al., 2010). However, it is unclear whether this modification is due to actual or subjectively perceived ambiguity in the signal. While these two possibilities are not necessarily mutually exclusive, the second hypothesis merits further discussion. The main argument here is that even before the onset of speech production, visual cues may provide information about nonnativeness of the speaker via facial cues, which may exaggerate the perceived foreignness of the

speaker. This, in turn, can affect how speech produced by nonnative speakers is processed, compared to that produced by native speakers.

5. Social Cues in Visual Speech

Even prior to the onset of speech, visual cues provide facial information that can indicate the nativeness of the speaker, regardless of its veracity. The objective of this section will be to provide credibility in the claim that this social aspect of the visual speech cues can affect its utility in enhancement of speech intelligibility. In this thesis we will focus on East Asian speakers. First, the literature on implicit social cognition will be reviewed. Then, the possible connection between implicit race-related associations and speech perception will be further elaborated upon.

In the realm of social psychology research, it has long been argued that implicit attitudes towards social markers exist, and that these can be dissociated from explicit attitudes and measured independently (Cunningham, Preacher, & Banaji, 2001; Greenwald & Banaji, 1995; Greenwald, McGhee, & Schwartz, 1998). In IAT, the participant is provided with two sets of visual stimuli. In the original experiment that had proposed its utility (Greenwald, et al., 1998), the first set was an object (flower or insect) and the second set words (pleasant or unpleasant). For each trial, the participant is presented with one of the stimuli, and asked to categorize it using one of the two response keys. In one experimental block, the participant may be asked to press the left key whenever an image of a flower is presented and the right key whenever an image of an insect is presented. In another experimental block, the same participant may be asked to press the left key for a pleasant word and the right key for an unpleasant word. These single category conditions comprise the practice phase. In the test phase, each trial in each block can be randomly pooled from either the object or the word stimuli set.

Correspondingly, the response mapping is also twofold. In the "congruous" condition, one response will be mapped to either flowers or pleasant words, while the other response will be mapped to either insects or unpleasant words. In the "incongruous" condition, however, one response will be mapped to either flowers or unpleasant words, while the other response will be mapped to either insects or pleasant words. It was predicted and subsequently confirmed that the participants with greater implicit negative attitude towards insects relative to flowers would be slower to respond in the incongruous condition relative to the congruous condition. The usefulness of this metric resides in the fact that such attitudes are arguably immune to conscious control, thereby alleviating the concern of social desirability bias (Crowne & Marlowe, 1960; Edwards, 1957; Fisher, 1993). This metric is called the implicit association test (IAT; Greenwald, et al., 1998).

Due to its simplicity and effectiveness, IAT has been extensively used in social psychology research. The method has been improved over the years (Greenwald, Nosek, & Banaji, 2003) and applied in a number of domains (for review, see Greenwald, Poehlman, Uhlmann, & Banaji, 2009). For instance, obesity research has revealed that obese individuals have more negative implicit attitudes towards high- vs. low-fat foods than do non-obese individuals, contrary to explicit food preference and intake patterns (Roefs & Jansen, 2002). Also, IAT has been demonstrated to be effective in predicting consumer choices (Maison, Greenwald, & Bruin, 2004). Furthermore, the claim that IAT tests automatic assumptions outside cognitive control has been strongly corroborated by the fact that participants are unable to "fake" their IAT scores even when explicitly instructed to do so (Banse, Seise, & Zerbes, 2001; Greenwald, et al., 2009).

These properties of IAT have especially advantageous in studying racial prejudices. Study of racial prejudice is often difficult due to the participants' desire to appear unprejudiced. Utilizing IAT, researchers have been able to tap into the domain of

implicit attitudes towards race and its relationship with explicit measures of racial prejudice (McConnell & Leibold, 2001). Further, these implicit-explicit links have been shown to be race-specific, such that negative implicit associations towards Turkish people predicted explicit prejudices towards Turkish people but not those towards East Asians, and vice versa (Gawronski, 2002). The alternative interpretation that could account for these findings is the lack of familiarity with the out-race group, but this notion has been discounted by other studies (Dasgupta, McGhee, Greenwald, & Banaji, 2000; Ottaway, Hayden, & Oakes, 2001), and the topic remains controversial (Kinoshita & Peek-O'Leary, 2005). Indeed, the interpretation of racial prejudice IAT results as the basis of accusation of racism is problematic, and the more balanced understanding is wanting. On the one hand, it has been found that explicit measures of racial attitudes predict self-perceived friendliness towards the members of the other race, while implicit measures predict how the said members and observers evaluate the participant's friendliness (Dovidio, Kawakami, & Gaertner, 2002). On the other hand, overtly harmful actions have been linked more robustly to implicit stereotypes rather than implicit attitudes (Rudman & Ashmore, 2007). Moreover, these IAT results regarding racial attitudes have been shown to be subject to modification through prejudice seminars (Rudman, Ashmore, & Gary, 2001), indicating that a considerable degree of plasticity exists in what is being measured by IAT. It has been suggested that the "prejudice" being measured is more reflective of shared cultural stereotypes which are not necessarily prejudiced (Arkes & Tetlock, 2004). The argument is that the IAT likely measures the extent to which each individual is exposed to an environment that endorses certain associations – some of which may be prejudiced – but that this measurement cannot be clearly dissociated from the implicit endorsement of these associations. The evidence behind this reasoning comes from findings that have shown that first, IAT and explicit

attitudes were not necessarily correlated, second, there were attitude-related behaviors predicted by explicit attitudes but not by IAT results, and third, exposure to new associations modified IAT results but not explicit attitudes (Karpinski & Hilton, 2001). Outside the IAT literature, it has been found that explicit rejection of the negative own-race stereotype can still negatively affect performance (stereotype threat; Steele, 1997). However, there are also attitude-related behaviors predicted by IAT results but not by explicit attitudes (Carney, Olson, Banaji, & Mendes, 2006). The distinction between shared stereotypes and "genuine" prejudice may be more semantic than scientific (Banaji, Nosek, & Greenwald, 2004). For the purposes of the present study, it is important to emphasize the current consensus that IAT results provide a window into processes that are elusive to conscious awareness. The current study does not elucidate the precise source of the observed preference towards racial associations.

There is evidence to suggest that East Asians are less likely to be automatically associated with "Americanness" in the United States of America (Devos & Banaji, 2005). In the study, the researchers had taken a multi-pronged approach to assessing how Americans define the American identity. The first step was to ask a large number of participants (N = 135) of their opinion on ethnic equality. A majority of the participants (88%) expressed the belief that Caucasian, African and Asian Americans should be treated equally. The rest of the participants (12%) expressed the belief that African Americans should be given priority. Secondly, the same participants were asked to report what constituted the set of core American values. It was revealed that among the values rated to be the most important was ethnic equality. However, when asked how much each ethnic group embodied the American values, ethnicity was found to be relevant. Specifically, Asian Americans were thought to be the least "American", Caucasian Americans the most American, and African Americans in between. Then, the researchers

had a separate set of Caucasian American participants (N = 28) explicitly report ethnicity-American associations and then complete an IAT in which participants were asked to associate three ethnicity pairs (Caucasian vs. African, African vs. Asian, Caucasian vs. Asian) with American or Foreign scenes. It was found that, explicitly, participants considered Caucasian and African ethnicities to be equally American, while the Asian ethnicity was considered to be less American to both ethnicities. However, the IAT results told a different story. Compared with the explicit ratings, participants displayed a larger tendency to consider the Caucasian ethnicity to be more American than the African ethnicity. Even greater was the tendency to associate the Asian ethnicity with Foreign identity. Participants did not differ in their comparison of embodiment of American concept between African and Asian ethnicities (Devos & Banaji, 2005). The main implications are threefold. First, holding abstract beliefs about ethnic equality in terms of rights or liberty does not guarantee that it would generalize to the realm of national identity. Second, explicit and implicit appraisals of Americanness across ethnicities differ. Third, among three ethnicities (Caucasian, African, and Asian), Asian Americans are the least likely to be automatically associated with American values.

The finding that Asian faces are less likely than Caucasian faces to be implicitly assumed to be native to the American environment is potentially relevant in nonnative speech perception by native listeners of English. Access to abstract information in facial cues does not require allocation of attention (Harry, Davis, & Kim, 2012). It can be hypothesized that a given listener will automatically assume an East Asian speaker to be nonnative to the American English speaking environment, even without the conscious intent of the listener to do so. Indeed, there is evidence to suggest that social cognition affects speech perception (for review, see Drager, 2010). For instance, acoustic boundaries between fricatives /s/ and /ʃ/ vary according to the sex of the speaker. When

ambiguous fricative is presented, listeners have been found to be affected by their knowledge of the sex of the speaker (Strand, 1999). Perception of vowels that are produced differently by speakers of different socioeconomic status (SES) has also been found to be affected by the visual information of the speaker manipulated to suggest higher or lower SES of the speaker (Hay, Warren, & Drager, 2006b). Similar results have been found with perception of ambiguous vowel production according to the available information regarding the speaker's nationality (Australia vs. New Zealand; Hay, Nolan, & Drager, 2006a; Canada vs. Michigan; Niedzielski, 1999). These effects of social cognition on speech perception have been found to override explicit knowledge. Even when participants could recognize the New Zealand mode of production in the presented vowel, they reported perception of Australian production when the word 'Australian' was displayed (Hay, et al., 2006a). Furthermore, the automaticity of the social information effect was generalized when the said information was not explicit but implicit, such as being exposed to stuffed toys that implicated nationality (e.g., kangaroos for Australia and kiwis for New Zealand) prior to speech presentation, without the listeners' knowledge of the role of the toys (Hay & Drager, 2010). From these findings, it can be hypothesized that native listeners of English, when presented with audiovisual speech produced by an East Asian speaker, will automatically assume that the speaker is a nonnative speaker of English. This social information, once processed, could significantly affect speech processing. In order to test this hypothesis, it will be necessary to demonstrate that a predictor of a listener's degree of association of East Asian faces and non-Americanness is linked to the listener's use of visual cues in speech perception.

6. PRESENT STUDY

In this study, I investigated the extent to which native listeners are able to utilize visual cues produced by nonnative speakers during speech processing. The basic design of the experiment was to present native listeners with speech produced by native or nonnative speakers, with or without visual cues. The speech stimuli were embedded in multi-talker babble noise (Van Engen et al., 2010) to induce a listening situation in which visual cues would benefit comprehension (Ross, et al., 2007; Sumby & Pollack, 1954). Adhering to this basic framework, it was first hypothesized that visual cues will benefit perception for both native and nonnative speech. However, there are competing hypotheses related to the relative contribution of visual cues to native and nonnative speech perception.

First, visual cues may benefit nonnative speech more than native speech. This prediction is based on findings that visual cues are more beneficial when there is room for improvement (Ross, et al., 2007; Sumby & Pollack, 1954), and that visual weighting in speech sound perception increases when the auditory signal is degraded (Fixmer & Hawkins, 1998; Nath & Beauchamp, 2011), such as is the case in nonnative speech sounds (Hazan, et al., 2010). Indeed, nonnative speech has been considered a form of an adverse listening condition (Mattys, et al., 2012). According to the principle of inverse effectiveness in audiovisual integration, visual cues are maximally effective when auditory cues are maximally degraded (Stein, Stanford, Ramachandran, Perrault Jr, & Rowland, 2009). The basis of this principle comes from the finding that the responses to multimodal stimuli by single neurons in the cat superior colliculus (SC) are inversely proportional to their responses to unimodal stimuli, which indicates that the neural response to audiovisual stimuli is not fixed but dynamic to the integrity of the auditory and visual streams (Alex Meredith & Stein, 1986). Recently, it has been found that the

BOLD activation in the human superior temporal cortex follows the principle of inverse effectiveness. Specifically, the cortical response to audiovisual speech stimuli were greater than what would have been additively predicted from the unimodal responses to auditory and visual cues, and this discrepancy increased as the overall SNR was decreased, reducing saliency of the stimuli (Stevenson & James, 2009). Applying this principle to nonnative speech perception where the auditory cues are degraded (Mattys, et al., 2012), it could be predicted that the magnitude of audiovisual integration will increase.

A second possibility is that visual cues may benefit native speech greater than non-native speech. This could be due to a number of reasons. Nonnative visemes are just as degraded as nonnative auditory cues. Nonnative visual cues may deviate from the target visemes, thus providing less advantage to speech intelligibility. Furthermore, degradation in the nonnative visual speech cues could lead to ineffective audiovisual integration. Importantly, as discussed before, visual information cues the non-native status of the speaker, which may exaggerate the perceived non-nativeness of the speaker. This prediction is supported by studies that show that speech perception is modulated by the extent of social information available to listeners (Drager, 2010). Indeed, there is a tendency for the East Asians to be perceived to be less likely to be native to America than are Caucasians (Devos & Banaji, 2005).. Indeed, abstract information in face stimuli can be processed preattentively (Harry, et al., 2012), which suggests that native listeners may automatically modify their perceptual patterns (McQueen, Norris, & Cutler, 2006) when facial cues suggest the nonnative status of the speaker (Devos & Banaji, 2005). Additional support for this prediction comes from EEG research on the temporal locus of audiovisual integration. It has been found that the amplitude of the auditory cortex response to speech sounds is reduced with the addition of visual cues, and that this effect

may happen as early as from 50 to 100 ms (van Wassenhove, Grant, & Poeppel, 2005). This finding that visual speech provides early predictive cues in auditory speech perception supports the prediction that the visual cues perceived to be indicative of the nonnative status of the speaker will interact with the processing of auditory cues. In order to dissociate the effects of sociophonetic variation in audiovisual nonnative speech perception from those of degraded signal in nonnative speech, an implicit association test (IAT; Greenwald, et al., 1998) can be administered to measure each participant's implicit bias. If the strength of the automatic Asian-Foreign associations was correlated with native-positive bias in speech intelligibility, and if this relationship only existed when visual cues exist and does not when they do not exist, then it could be argued that social cognition had affected nonnative speech perception. Additionally, explicit ratings of accentedness levels in speech perception by an independent group of participants could provide additional evidence towards the effect of visual cues on the perception of nonnativeness (Smiljanić & Bradlow, 2011).

In the current study, participants were presented with sentences produced by native English and native Korean (nonnative in English language) speakers, with or without visual cues. The sentences were presented mixed with six-talker babble. Participants were asked to transcribe the sentences. The accuracy of the keywords in each sentence was calculated and compared across the four conditions: native audiovisual (AV), native audio-only (AO), nonnative AV, and nonnative AO. The effects of nativeness and modality (AV vs. AO) are studied, as well as the interaction between the two factors. The pattern of speech intelligibility enhancement in AV or AO condition is correlated against each participant's IAT score of Asian-Foreign association. Additionally, foreign accent ratings for the speech stimuli in all four conditions are obtained from independent participants to ascertain that the nonnative speech stimuli are

indeed perceived to be accented and to discover, if any, effects of visual cues on the perception of foreign-accented speech. Finally, the participants were also presented with McGurk stimuli produced by native and nonnative speakers. This was performed to confirm a previous finding of increased visual weighting for nonnative speech stimuli (Hazan, et al., 2010) in our sample.

Chapter 2: Methods

1. PARTICIPANTS

Young adults (N = 27; 18 female; ages: 18 to 39) were recruited from the University of Texas community and received monetary compensation or research credit for their participation. All participants were monolingual native American English speakers with no language problems. All participants passed a hearing screening (audiological thresholds < 25 dB HL across 0.5, 1, 2, and 4 kHz). Six of the twenty-seven participants (3 female) provided accent ratings. The remaining twenty-one listeners (15 female) participated in the speech perception in noise (SPIN), McGurk perception, and implicit association test (IAT) tasks. Participants did not overlap between the accent rating and SPIN task.

2. MATERIALS

2.1 Audiovisual Speech Stimuli

Four native American English (2 female) and four native Korean speakers (2 female) produced eighty sentences with four keywords (e.g., "The GIRL LOVED the SWEET COFFEE."; Calandruccio & Smiljanic, 2012; Figure 1a) and CV syllables with three voiced and unvoiced stop consonants (bilabial: /ba/, /pa/; alveolar: /da/, /ta/; velar: /ga/, /ka/). Each syllable was repeated three times. The video track was recorded using a Sony PMW-EX3 studio camera, and the audio track was recorded with an Audio Technica AT835b shotgun microphone placed on a floor stand in front of the speaker. Camera output was processed through a Ross crosspoint video switcher and recorded on an AJA Pro video recorder. The session was conducted on a sound-attenuated sound stage at The University of Texas at Austin.

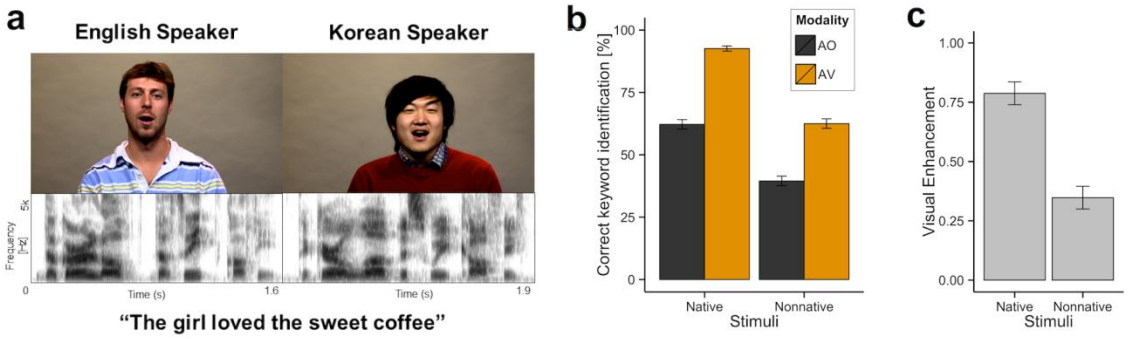

Figure 1: (A) Visual (upper panel) and auditory (lower panel) streams of the sentence "The girl loved the sweet coffee" produced by native and nonnative speakers. (B) Percentage of the keywords correctly identified for the speech perception in noise task for native English and Korean speakers, without and with visual cues. (C) Visual enhancement measures compared between native English and Korean speakers.

2.2 SPIN Masker

Six native speakers of American English (3 female) produced thirty simple, meaningful sentences (Bradlow & Alexander, 2007; Van Engen, et al., 2010) were used for the 6-talker babble track used as the masker. All sentences were RMS amplitude normalized in Praat (68 dB; Boersma & Weenink, 2010), concatenated and mixed across all six talkers in Audacity (Audacity Developer Team, 2008), trimmed to 50 s and RMS amplitude normalized in Praat (69 dB; Boersma & Weenink, 2010) to yield in an signal to noise ratio (SNR) of -4 dB. Forty unique random samples of this continuous masker stream were mixed with the target sentences using Adobe Audition (Riley, 2008), so that a 500 ms stream of auditory noise and a freeze frame of the video enveloped the onset and the offset of the target sentences. The stimuli were encoded into a DV Video (dvsd) stream of 720x576 resolution and 30 fps, with an uncompressed mono PCM S16 LE (araw) audio stream of 22,050 Hz sample rate, 16 bps, and 352 kb/s bitrate. All video editing was conducted in Final Cut Pro (Weynand, 2010).

2.3 IAT Stimuli

Ten young adult Asian (5 female) and ten Caucasian (5 female) face images were used for Caucasian vs. Asian face categories (Minear & Park, 2004). All face images had been edited to exclude hair, face contour, ear, and neck information, then rendered into grayscale with constant luminosity (Goh, Suzuki, & Park, 2010). Public domain images of ten iconic American scenes (Grand Canyon, Statue of Liberty, Wrigley Stadium, Golden Gate Bridge, Pentagon, Liberty Bell, White House, Capitol, New York Central Park, Empire State Building) and ten non-American foreign scenes (Eiffel Tower, Pyramids, Angkor Wat, London Bridge, Brandenburg Gate, Stonehenge, Great Wall of China, Leaning Tower of Pisa, Sydney opera House, Taj Mahal) were obtained online and used for American vs. Foreign scene categories. No scene image contained face information. All images were cropped to a square proportion (Figure 2a).

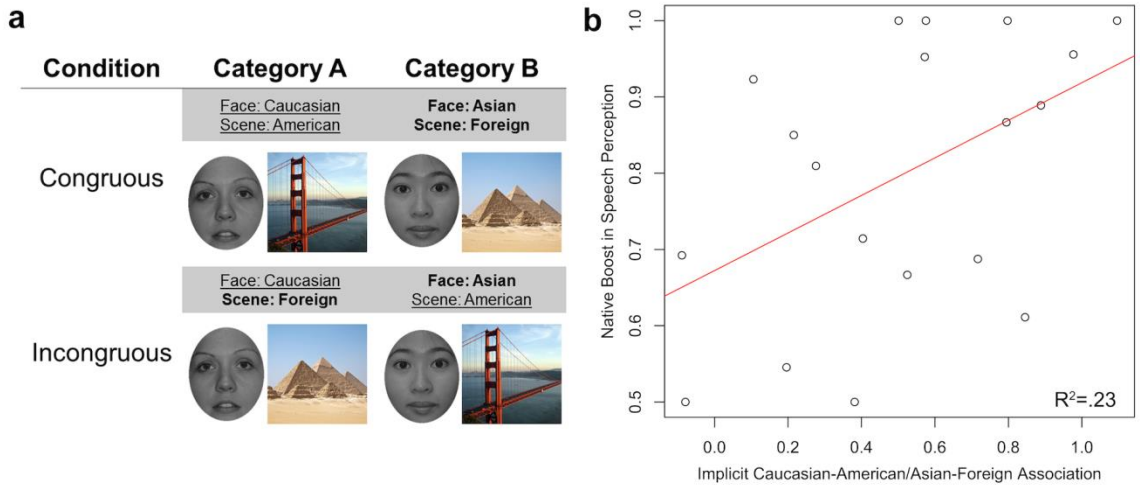

Figure 2: Implicit association test. (A) Face (10 Caucasian; 10 Asian) and scene (10 American; 10 Foreign). In the congruous condition, participants were instructed to group Caucasian faces and American scenes together, and Asian faces and foreign scenes together. In the incongruous condition, participants were instructed to group Caucasian faces and foreign scenes together, and Asian faces and American scenes together. (B) IAT scores and the native boost when visual cues were available positively correlated with each other, $r(17) = .482$, $p = .037$, $R^2 = .23$.

2.4 McGurk Syllables

From each speaker, the experimenters chose the most representative production of each syllable, excluding samples with extraneous facial movements unrelated to speech production, and equating prosody and duration across the six syllables. The video was segmented via Final Cut Pro (Weynand, 2010) to include the entirety of initiation and termination of both visual and auditory syllable production. The audio and video segments were then intermixed within each talker and voicedness of the consonants to produce the following types of stimuli: (a) McGurk-incongruent stimuli (MIS) that contained auditory bilabial (e.g., /ba/) and visual velar (e.g., /ga/) consonants; (b) Non-McGurk-incongruent stimuli (NMIS) that included all five syllable combinations that were incongruent (i.e., the auditory and visual consonants were different) but did not include the combination (a); and (c) congruent stimuli (CS), which were not modified and had same visual and auditory information (Figure 3a). The audio of each video file was then extracted in the loseless mono PCM S24 LE (araw) format at the sample rate of 48 kHz with 24 bits per sample and RMS normalized to 72 dB using the Praat software (Boersma & Weenink, 2010), and remixed with the original video files. All video files were exported using DV Video (dvsd) codec with the resolution of 720x576 and frame rate of 29.97 frames per second, while the audio stream was exported using PCM S16 LE (araw) codec mixed down to a mono channel at the sample rate of 48 kHz with 16 bits per sample. The four speakers' production of voiced and voiceless syllables and the intermixing process yielded in eight MIS, 40 NMIS, and 24 CS.

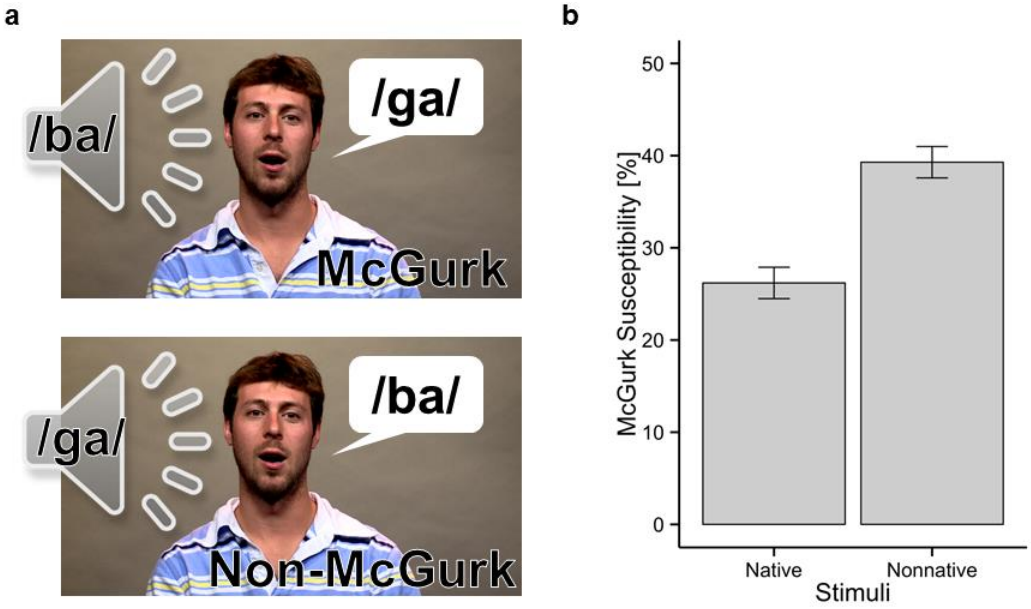

Figure 3: McGurk Effect. (A) McGurk (upper panel; auditory /ba/; visual /ga/) and non-McGurk (lower panel; auditory /ga/; visual /ba/) stimuli. (B) McGurk susceptibility was higher for native stimuli than for nonnative stimuli.

3. PROCEDURES

3.1 Speech Perception in Noise

Participants were placed in a sound-attenuated booth. Forty speech-in-noise stimuli were presented in a randomized order. Each sentence was randomly produced exclusively by only one of the four speakers, in either audio-only (AO) or audiovisual (AV) condition. In the AO condition, the video track was replaced with a fixation cross. After each stimulus presentation, participants were instructed to type the response using a computer keyboard. The responses were then scored for by-keyword accuracy. Spelling mistakes or homophones were also scored to be correct.

3.2 Accent Rating

Participants were placed in a sound-attenuated booth and listened to the forty sentences produced by all four speakers in both AO and AV conditions, yielding a total

of 320 stimuli. In the AO condition, the video track was replaced with a fixation cross. The presentation order was randomized. Participants were instructed to rate how accented each sentence was on a 1-to-9 Likert scale: 1 = no foreign accent; 9 = very strong foreign accent. This scale had been used in a previous study which had examined the relationship between speech intelligibility and the perceived accentedness (Smiljanić & Bradlow, 2011).

3.3 IAT

Participants were instructed to perform a response time task in which they were to respond as quickly as possible without sacrificing accuracy. Participants were not informed of the objective of the experiment. For each trial, a face or scene stimulus was displayed on the screen. In the congruous category condition, participants had to press a key on the keyboard when they saw a Caucasian face or an American scene, and another key for an Asian face or a Foreign scene. In the incongruous category condition, participants had to press a key for a Caucasian face or a Foreign scene, and another key for an Asian face or an American scene. Each condition was presented twice with the key designations switched in a randomized order. In all trials, incorrect responses led to the corrective feedback of "Error!".

3.4 McGurk Effect

The participants were seated in front of a computer monitor. The stimuli were presented once in a randomized sequence. The participants were instructed to report the syllable heard (chosen from BA, DA, GA, PA, TA, and KA), and then provide a confidence rating on each answer on a Likert scale of 1 to 7, with 1 indicating "Not sure," 7: "Absolutely sure", and 4: "Somewhat sure." The presentation of successive stimuli

were self-paced initiated by a button press, followed by a display of a fixation cross of 500 ms duration preceding video presentation.

4. Data Analysis

4.1 Speech Perception in Noise

In analyzing the SPIN data, the linear mixed model was used instead of ANOVA. This was done because the linear mixed model allows analysis of data from individual trials instead of averaging across conditions. Additionally, the approach allows the researcher to include random effects modeling that would account for the variability arising from individual participants, items, etc. (Baayen, Davidson, & Bates, 2008).

SPIN outcome (correct vs. incorrect) for each word response was entered as the dependent variable through the linear mixed model using a binomial logit link (Bates, Maechler, & Bolker, 2012). Fixed effects included modality and nativeness and their interaction term, corrected for random by-subject, by-sentence, and by-word intercepts.

Visual benefit and native-speaker benefit were also measured using the following equations: AV boost = (AV – AO) / (1 – AO); native boost = (Native – Nonnative) / (1 – Nonnative). These equations follow established method of calculating enhancement from additional cues (Sommers, Tye-Murray, & Spehar, 2005).

4.2 Accent Ratings

Accent rating scores provided by the participants (n = 6) were converted to continuous percentage scale of native-like accent: 0%: least native-like; 100%: most native-like. For instance, a rating of 9 (very strong foreign accent) would be converted to 0%, while a rating of 1 (no foreign accent) would be converted to 100%. A linear mixed effects analysis (Bates, et al., 2012) was run on these percentage ratings as the dependent variable. Fixed effects were modality condition (AV vs. AO), nativeness of the speaker

(English vs. Korean), and their interaction term, corrected for the random by-subject and by-sentence intercepts. P-values of the fixed effects were calculated with Markov Chain Monte Carlo sampling (n = 10000).

4.3 IAT

Response times (RT) were scored to yield one IAT score per participant. The scoring algorithm compared the RT differences across congruous (Caucasian-American and Asian-Foreign) vs. incongruous (Caucasian-Foreign and Asian-American) conditions, while penalizing for incorrect responses and excluding from the analysis artifact trials with extreme RTs (Greenwald, et al., 2003). A higher IAT score indicated a greater implicit bias towards making Caucasian-American and Asian-Foreign associations (Devos & Banaji, 2005). An outlier analysis was performed (< ±1.5*SD; n=19). The IAT scores were regressed against SPIN native boost scores for AV and AO conditions separately, using Pearson's product-moment correlational analysis.

Linear mixed effects analyses (Bates, et al., 2012) were run with RT in milliseconds as the dependent variable. In the first analysis, only the category condition (congruous vs. incongruous) was entered as the fixed effect to ascertain the overall phenomenon of implicit association. In the second analysis, the fixed effects were category condition and SPIN native boost scores (AV) for each participant. By-subject random intercepts were included. P-values of the fixed effects were calculated with Markov Chain Monte Carlo sampling (n = 10000).

4.4 McGurk Effect

Each participant's response for the native and nonnative McGurk-Incongruent-Stimuli (MIS) were separately coded into visual (/ga/ or /ka/), auditory (/ba/ or /pa/) or fused (/da/ or /ta/) percepts. The percentage of the fused percepts out of four stimuli in

each condition was used as the measure of susceptibility to the McGurk effect. The McGurk measures for native and nonnative speakers were compared using a paired t-test.

An additional analysis was performed with the confidence ratings provided by the participants for each response was used to weigh the perceptual responses. The ratings (7: Absolutely sure; 1: Not sure) were converted onto a linear 0 to 1 scale denoting the participants' confidence of the perceptual experience, and used to weigh the raw proportions of fused percepts. This was done to ascertain that the nativeness effect on McGurk susceptibility, if any, will persevere when the listeners' judgment of ambiguity was taken into consideration.

Chapter 3: Results

1. Speech Perception in Noise

62.4% of the keywords produced by native English speakers were identified correctly by the participants in AO, and 92.9% in AV. 39.5% of the keywords produced by native Korean speakers were identified correctly by the participants in AO and 62.5% in AV (Figure 1b). Comparison of AV/AO ratios based on raw values suggests a 48% increase for English and 58% for Korean speakers. However, comparing simple ratios to calculate the enhancement biases against conditions with higher reference score, in this case the native English speaker condition. Since, the percentage values represent the average probability that each word in a sentence will be perceived correctly in a given condition, the null distribution follows the binomial distribution of "correct" or "incorrect". As performance reaches the positive extreme, the null probability associated with performance exponentially decreases, making it more difficult for the listener to improve the same numeric amount in performance. Therefore, a linear comparison of simple ratios of percentage scores is inadequate. The analytic method must take into account the exponentially increasing difficulty for higher reference (AO) scores.

Traditionally, this objective has been achieved by calculating a "visual enhancement" score where the (AV – AO) difference is corrected by the denominator (1 – AO). Hence, the visual enhancement is positively adjusted for higher AO scores, and negatively for lower AO scores (Grant & Seitz, 2000; Sommers, et al., 2005). The visual enhancement for native speech ($M = .79$; $SD = .18$) was higher than for nonnative speech ($M = .35$; $SD = .32$), $t(20) = 6.49$, $p < .0001$, indicating that the visual cues benefit native speech more than nonnative speech (Figure 1c).

A more direct approach would be to implement the generalized linear mixed effects analysis which estimates the effect of modality and nativeness conditions on the

logit probability that a given word will be perceived correctly (Bates, et al., 2012). Four estimates are provided: (a) the intercept; (b) effect of AV relative to AO; (c) effect of Korean speakers relative to English speakers; (d) AV-Korean interaction. The interaction term is of main interest in analyzing this study. A positive interaction term would indicate that the AV modality benefits nonnative speech more than native speech, where a negative interaction term would indicate the opposite, that the AV modality benefits native speech more than nonnative speech. In the mixed effects analysis, the intercept was significant, $b = 0.6951$, $SE = .2621$, $z = 2.65$, $p = .008$. The nativeness effect was significant, $b = -1.1925$, $SE = .1134$, $z = -10.51$, $p < .0001$, such that word recognition in noise was better for English speakers than for Korean speakers. The modality condition effect was significant, $b = 2.1767$, $SE = .1624$, $z = 13.40$, $p < .0001$, such that keywords were more correctly identified in AV than in AO. The nativeness by condition interaction effect was significant, $b = -.1.1088$, $SE = .1974$, $z = -5.62$, $p < .0001$, such that the AV benefit was greater for English than for Korean speakers. The AV nonnative estimate would have been 84.3% without the interaction term; it is 63.9% with the interaction term. This finding indicates reduced efficiency in audiovisual integration for perception of nonnative speech relative to native speech.

2. ACCENT RATING

The average native-like rating for native English speakers was 96.2% in the audio-only condition (AO) and 97.1% in the audiovisual condition (AV). The average native-like rating for native Korean speakers was 20.7% in AO and 18.9% in AV, exhibiting an opposite pattern due to visual cues from that for the native English speakers. The lmer analysis revealed that the intercept was significant, $b = 96.1725$, $SE = 2.0581$, $t = 46.73$, $p < .0001$. The nativeness effect was significant, $b = -75.4605$, $SE =$

.8258, t = -91.38, p < .0001, with Korean speakers rated as more foreign-accented. The modality condition effect was not significant, b = .9324, SE = .8258, t = 1.13, p = .2630, indicating that the inclusion of visual cues did not have an overall effect on the perception nativeness. However, the nativeness by modality condition interaction effect was significant, b = -2.7173, SE = 1.1675, t = -2.33, p = .0166, an effect explained by the numerical observation that the native Korean speakers were rated as more foreign-accented in AV relative to AO, while the native English speakers were rated to be less accented in AV than in AO.

3. IAT

IAT scores, as a whole, were significantly higher than zero (M = .511; SD = .347), t(18) = 6.41, p < .0001, indicating participants had an overall bias toward congruous associations (Caucasian-American and Asian-Foreign). IAT scores not significantly different from zero would have indicated that there was no overall pattern of bias consistently observed for all participants. IAT scores significantly lower than zero would have indicated that the participants had an overall bias toward the incongruous associations (Caucasian-Foreign and Asian-American). IAT scores were positively correlated with the native boost in AV, r(17) = .482, p = .037, indicating that participants with higher tendency to make an implicit Caucasian-American and Asian-Foreign association were more likely to show enhanced performance for native than for nonnative sentences in AV (Figure 2b). In contrast, IAT scores were not significantly correlated with native boost in AO, r(17) = .064, p = .80, indicating that the bias against incongruous associations was not related to relative performance across sentences produced by English and Korean speakers in AO. In other words, a consistent

relationship between IAT scores and the native boost in the SPIN task only existed when the visual cues were available for the listeners.

Next, the linear mixed effects analyses were conducted to directly assess the impact of the metrics on the response times. First, the model with only the category condition as the fixed effect was run to ascertain that the task had functioned as originally intended. The intercept was significant, $b = 858.33$, $SE = 41.77$, $t = 20.55$, $p < .0001$. The incongruous condition showed a significant effect, $b = 174.11$, $SE = 15.86$, $t = 10.97$, $p < .0001$, indicating that the responses in the incongruous condition were significantly slower than in the congruous condition by approximately 174 ms. Second, the model with category condition, SPIN native boost scores (AV), and their interaction term was run. The intercept was significant, $b = 1059.77$, $SE = 200.94$, $t = 5.27$, $p < .0001$. The incongruous condition effect was not significant, $b = -44.07$, $SE = 76.66$, $t = -.58$, $p = .58$, nor was the SPIN native boost effect, $b = -254.92$, $SE = 248.81$, $t = -1.03$, $p = .31$. However, there was a significant interaction between the incongruous condition and the SPIN native boost scores, $b = 276.10$, $SE = 94.92$, $t = 2.91$, $p = .0024$. The participants with higher SPIN native boost scores were also likely to respond slower to incongruous stimuli, which indicates that the participants with higher degree of bias towards making Caucasian-American and Asian-Foreign assumptions were more likely to process native AV speech better than nonnative AV speech. The fact that the incongruous condition effect was no longer significant with the inclusion of the SPIN native boost and interaction terms suggest that the same underlying procedure gave rise to both the IAT effects and enhancement of native speaker intelligibility (or conversely, disruption of nonnative speaker intelligibility) when visual cues are available.

4. McGurk Effect

The average McGurk susceptibility was 26% (SD = 27%) for syllables produced by native speakers, and 39% (SD = 36%) for those produced by nonnative speakers. The difference was significant, $t(20) = 3.99$, $p = .00072$, where participants were more likely to report an audiovisually fused percept for the speech stimuli produced by nonnative speakers than for those by native speakers (Figure 3b).

The weighted average McGurk susceptibility was 15% (SD = 17%) for syllables produced by native speakers, and 23% (SD = 22%) for those produced by nonnative speakers. The difference was significant, $t(20) = 2.83$, $p = .01038$, where participants were more likely to report an audiovisually fused percept for the speech stimuli produced by nonnative speakers than for those by native speakers, when these percepts were weighted with the confidence ratings.

Chapter 4: Discussion

1. Summary of Results

The goal of this study was to examine the role of audiovisual integration in perception of speech produced by nonnative speakers. Native listeners of English were instructed to listen to sentences presented in noise and report the words that they had perceived. The sentences were produced by native and nonnative speakers of English, with or without visual cues. It was hypothesized that native speech stimuli would yield greater accuracy in word identification that would nonnative speech stimuli. Also, it was hypothesized that visual cues would help word identification overall. However, competing hypotheses existed concerning whether nonnative visual cues would have an enhanced or diminished role in improving speech intelligibility in noise.

In line with the initial predictions, it was confirmed that native speech perception was easier for the listeners than was nonnative speech perception. This effect of speaker nativeness on speech perception in noise is in line with previous findings (Munro, 1998; Munro & Derwing, 1995a, 1995b). Also, it was confirmed that the availability of visual cues enhance speech intelligibility regardless of the nativeness of the speaker. However, perception of nonnative accent was subtly affected by the presence of visual cues in a different manner. Visual cues had differential effects on the perceived accentedness of the native and nonnative speakers, where inclusion of visual cues led native speakers to be perceived as less accented and nonnative speakers to be perceived as more accented. This finding is in accordance with the previous findings that listeners incorporate visual cues to improve speech perception in a compromised listening environment (Erber, 1975; Girin, et al., 2001; Grant & Seitz, 2000; MacLeod & Summerfield, 1987, 1990; Ross, et al., 2007; Sumby & Pollack, 1954). Finally, audiovisual integration was found to be less effective in resolving nonnative speech in noise in comparison to native speech. This

effect was corroborated via linear mixed effects analysis and comparison of visual enhancement scores. Although the audiovisual (AV) condition yielded more accurate perception for both native and nonnative speech than did the audio-only (AO) condition, this effect was more pronounced for native speech than for nonnative speech.

Furthermore, the phenomenon of increased visual weighting for nonnative syllables was replicated (Hazan, et al., 2010). Participants were more likely to report fused percepts for audiovisually incongruent McGurk syllables (McGurk & MacDonald, 1976).

2. REDUCED BENEFIT FROM AUDIOVISUAL INTEGRATION FOR NONNATIVE SPEECH

It has been well established that visual cues aid speech perception in noise (Erber, 1975; Girin, et al., 2001; Grant & Seitz, 2000; MacLeod & Summerfield, 1987, 1990; Ross, et al., 2007; Sumby & Pollack, 1954). However, it had been unclear whether this effect also holds for the perception of speech produced by nonnative speakers (Hazan, et al., 2010). The present study was designed to address this question by having the listeners process native and nonnative speech in noise with and without visual cues. Two opposing predictions had been proposed regarding the relative efficiency of audiovisual integration in native vs. nonnative speech perception in noise.

The first hypothesis stated that the perception of nonnative speech would benefit more from the availability of visual cues, since listeners tend to place greater weighting on the visual stream of speech when the auditory stream is more degraded, as is the case in nonnative speech (Fixmer & Hawkins, 1998; Nath & Beauchamp, 2011; Sumby & Pollack, 1954). Indeed, it had been previously found that native listeners are more likely to rely on visual cues when resolving phonemic-level ambiguity in speech sound stimuli for nonnative speech than for native speech (Hazan, et al., 2010). If visual cues are

beneficial for speech perception in noise and the listeners are more likely to incorporate visual cues for nonnative speech sounds, it would logically follow that audiovisual integration would be more beneficial for speech intelligibility when the listeners listen to nonnative speakers. This prediction, as has been demonstrated, was not realized in the experiment. On the contrary, although visual cues increased speech intelligibility for nonnative speakers, the visual enhancement was lower than that for native speakers. In order to reconcile the sentence-related audiovisual integration results with the seemingly opposite findings from the current syllable-related visual weighting results and the previous syllable perception study (Hazan, et al., 2010), a few interpretations could be offered. First, it is possible that although the listeners weighed the visual cues more heavily for nonnative speakers, the cues that they had received were too degraded relative to native visual cues. Second, it is possible that although the listeners weighted the visual cues more heavily for nonnative speakers and these cues had comparable signal integrity as those of native speakers, an additional factor prevented beneficial audiovisual integration. The second hypothesis better takes into account this additional factor.

The second hypothesis, in contrast to the first, had been that the perception of native speech would benefit more from visual cues than would that of nonnative speech. This prediction had arose from the literature in race cognition research that suggests East Asian faces are more likely to be perceived to be foreign to the U.S. (Devos & Banaji, 2005), that abstract facial information processing can be preattentive (Harry, et al., 2012), and that social information can affect patterns in speech perception (Drager, 2010), due to the dynamic nature of the process (McQueen, et al., 2006). The results from this experiment indicated reduced effectiveness in audiovisual integration for nonnative speech perception in noise. In order to dissociate the simple account of signal-driven inefficiency in visual cues from the more complex sociophonetic interpretation (Drager,

2010), the native speech advantage in the AV condition was regressed against each participant's Asian-Foreign IAT score (Devos & Banaji, 2005). It was found that the more a participant was likely to associate East Asian faces with foreignness, the greater disparity in speech intelligibility between native and nonnative stimuli. In other words, availability of visual cues induced the participants to be more effective for resolving native speech in noise than for nonnative speech, and this disparity was proportional to the participants' tendency to automatically assume that East Asian faces are of foreign origin. Moreover, this relationship was not found when visual cues were absent.

Hence, a significant degree of individual variability exists in the ability to incorporate visual cues in nonnative speech, and a portion of this variability is attributable to an implicit association between East Asian speakers and nonnative status in the American English language environment. It is argued that while audiovisual integration is beneficial for nonnative speech perception in noise, its efficiency is compromised due to the social cognition of native listeners (Drager, 2010). In other words, social perception of nonnative status of the speakers accounts for at least a portion of the variability in the ability to use visual cues in nonnative speech. While it could still be argued that the IAT results may simply reflect familiarity with and exposure to East Asian speakers (per Arkes & Tetlock, 2004; c.f., Dasgupta, et al., 2000; Quillian, 2008), this position does not explain why such a relationship should be absent in the AO condition. If the IAT in the present study reflected experience with East Asian speakers, then a similar, albeit arguably smaller, relationship should have been observed even when visual cues were not present from nonnative speakers. Instead, only 0.4% of the variance in the native boost in AO was explained by the variance in IAT, and it is unlikely that this lack of effect can be attributed to low power, given that the variance in IAT explained an incomparably higher proportion of 23% of the variance in native boost in SPIN when

visual cues were available. Although the results do not indicate that the speaker-identity implicit association is the sole source of reduced audiovisual integration efficiency for nonnative speech perception, they strongly suggest the existence of sociophonetic mediators of native vs. nonnative speech perception.

Returning to the apparent discrepancy between the native vs. nonnative audiovisual integration patterns across syllables and sentences (Hazan, et al., 2010), the following conclusion could be drawn tentatively. There was a positive correlation between the degree of each participant's implicit social bias (Asian-Foreign association) and enhanced native speech intelligibility relative to nonnative speakers. Therefore, social cognition partially hinders beneficial audiovisual integration, despite increased visual weighting for nonnative speakers. Of course, it cannot be overlooked that the IAT measure only predicts a portion of the variance in the native boost – or nonnative degradation – in audiovisual speech intelligibility. Given the significantly reduced intelligibility for audio-only nonnative speech relative to audio-only native speech, it is more than reasonable to assume that the rest of the variance unexplained by the variance in implicit social cognition should be attributable to simple signal-driven degradation in the speech stimuli produced by nonnative speakers.

It is again emphasized that the results from the present study do not indicate a relationship between racism and speech perception. There is still an ongoing debate regarding - the extent to which race cognition IAT truly measures prejudice (Arkes & Tetlock, 2004), and the IAT administered in this experiment does not deal with positive or negative stereotypes associated with race (Devos & Banaji, 2005). The results from the accent ratings provided by an independent set of participants also support this claim, since participants were more likely to consider nonnative speech to be more accented when visual cues were available, although the auditory signal had been identical. It

appears that visual cues not only affect speech perception but also the perception of speakers, and the degree of this effect is considerably variable across different listeners.

3. IMPLICATIONS

Currently, the effort to enhance the intelligibility of speech produced by nonnative speakers of English in the United States is focused on "reduction" of foreign-accents by training these speakers to sound more like native speakers (Jokisch, Koloska, Hirschfeld, & Hoffmann, 2005; Rosini, 1997; Seferoğlu, 2005). However, these accent reduction programs are often ineffectual in meeting their objectives of having nonnative speakers sound like native speakers. Moreover, the presence of foreign-accented speech is not directly linked to diminished intelligibility (Derwing & Munro, 2009). Furthermore, the current accent reduction paradigm is burdening nonnative speakers with an increased demand on their speech output, when they already have low proficiency in the target language.

In the present study, listeners with non-linguistic social bias have been found to be more inefficient in utilizing visual cues for nonnative speech processing. Regardless of whether the IAT reflects familiarity with a particular subset of nonnative speakers (Arkes & Tetlock, 2004) or a form of genuine implicit social cognition (Devos & Banaji, 2005), it stands to reason that both are modifiable. On the one hand, listeners without much experience with nonnative speakers could be exposed to more instances of nonnative speech. Indeed, it has been demonstrated that not only can native speakers be trained to process a specific nonnative speech style better (Bradlow & Bent, 2008), but that extensive training sessions with multiple nonnative speech styles allow generalization of the training benefits to a novel nonnative accent (Baese-Berk, Bradlow, & Wright, 2013). On the other hand, it has also been reported that implicit social associations are subject to

modification through goal-directed training (Rudman, et al., 2001), from which, according to the present findings, it can be conjectured that social cognition training to reduce the implicit Asian-Foreign association may increase the efficiency in audiovisual integration for East Asian nonnative English speech. These studies altogether suggest that not all of the reduced intelligibility in nonnative speech is signal-driven, but that room for improvement exists on the listener's part.

The findings from these studies (Yi et al., 2013; Baese-Berk, et al., 2013; Bradlow & Bent, 2008) provide an important insight into how the problem of nonnative speech should be approached through addressing listener-related effects. Not only can the native listeners be trained to attain greater levels of expertise in nonnative speech perception (Baese-Berk, et al., 2013; Bradlow & Bent, 2008), but a social cognition modification plan could be implemented to reduce the native listeners' implicit Asian-Foreign association that hinder efficient audiovisual nonnative speech processing (Yi et al., 2013; Rudman, et al., 2001). However, the possibility of training benefits does not necessarily indicate that nonnative speech is not degraded. Although clearly lacking rich cues that natural speech offers, artificially manipulated speech stimuli such as vocoded speech or sine wave speech also allow room for improvement following extensive training (Davis, et al., 2005; Sheffert, Pisoni, Fellowes, & Remez, 2002). It is difficult to dissociate the effects of native listeners' familiarity with the native speaking style from those of the inherent perturbation of speech processing caused by nonnative speech (Floccia, Butler, Goslin, & Ellis, 2009; Floccia, Goslin, Girard, & Konopczynski, 2006). Moreover, the current experiment only presents results for native Korean speakers, who are of East Asian descent. In order to remedy these limitations, further studies are necessary.

4. FUTURE DIRECTIONS

There are a number of ways to further dissociate the factors of speaker-driven signal degradation vs. listener-driven experience in nonnative speech perception. The first is to recruit nonnative listeners of English for an identical SPIN paradigm. Nonnative listeners of English have low exposure to both native and nonnative speaking styles of English than do native listeners. If the signal degradation in the nonnative speech stimuli is the dominant factor behind reduced efficiency in audiovisual integration, then nonnative listeners will exhibit a similar enhancement for native speech as for native listeners. However, if the listener-driven experience is the dominant factor, then nonnative listeners would be expected to exhibit less of a native speech enhancement.

The second way to dissociate the effects of social cognition from speech signal degradation is to subject native listeners to static vs. dynamic conditions of audiovisual speech produced by native and nonnative speakers. In the static audiovisual condition, visual cues are present only so far as to reveal the speaker identity and therefore hint at the native status of the speaker via listeners' implicit Caucasian-American and Asian-Foreign associations. However, since in the static audiovisual condition the speech articulators will remain steady in a freeze frame or be obstructed from view by a visual masker, signal degradation in the visual cues will not contribute to the modification in speech intelligibility, if any. In this case, performance discrepancy across static vs. dynamic audiovisual conditions can be attributed to the aspect of audiovisual integration pertaining to speech cues only, while comparing the effect of static visual cues on native vs. nonnative speech will be informative of the extent of the effect of non-speech social cues.

The third way to dissociate the social factors from speech factors is to present speech produced by two additional subgroups of English speakers. The first group will be

nonnative speakers of English with Caucasian appearance, for whom the social cognition driven by the face information will not hinder audiovisual speech integration, but the irregularities in the nonnative speech production will do so. The second group will be native speakers of English with Asian appearance, for whom the social cognition may interfere with optimal audiovisual integration, but the speech signal will not have been degraded due to the nonnative status of the speakers. Therefore, this study will be a two-by-two design where the two factors are appearance (Caucasian vs. Asian) and nativeness (native vs. nonnative). This simple experiment is expected to improve our understanding of the complex effects of listener-driven and speaker-driven factors behind nonnative speech perception.

5. CONCLUSIONS

Visual cues help listeners understand speech better in more challenging listening situations (Ross, et al., 2007; Sumby & Pollack, 1954). Nonnative speakers produce speech that is more difficult to understand than is speech produced by native speakers, especially in noise (Munro, 1998; Munro & Derwing, 1995a, 1995b). This study examined the extent of audiovisual integration in the perception of nonnative speech produced by native Korean speakers in noise. It was revealed that the native listeners of English are not as adept at using visual cues to enhance the intelligibility of nonnative speech. Moreover, the extent of relative nonnative degradation in audiovisual speech intelligibility was linked to the listeners' implicit social cognition of Caucasian-American and Asian-Foreign associations (Devos & Banaji, 2005). It is argued from these results that non-speech social cognition plays a significant role in nonnative speech perception, and therefore listener-driven social and speech perceptual modification strategies should

be considered in improving the intelligibility of nonnative speech (Baese-Berk, et al., 2013; Bradlow & Bent, 2008; Rudman, et al., 2001).

References

Alex Meredith, M., & Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Research*, *365(2),* 350-354.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "would Jesse Jackson 'fail' the implicit association test?". *Psychological Inquiry*, *15(4),* 257-278.

Audacity Developer Team. (2008). Audacity (Version 1.2.6) [Computer software]. Available: http://audacity.sourceforge.net/download.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59(4),* 390-412.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America, 133(3),* EL174-EL180.

Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2004). No place for nostalgia in science: A response to Arkes and Tetlock. *Psychological Inquiry, 15(4),* 279-310.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie, 48(2),* 145-160.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes.

Benguerel, A.-P., & Pichora-Fuller, M. K. (1982). Coarticulation effects in lipreading. *Journal of Speech, Language and Hearing Research, 25(4),* 600.

Benoit, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech, Language and Hearing Research, 37(5),* 1195.

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer [Computer program], Version 5.1. 44.

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America, 121,* 2339.

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America, 112,* 272.

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

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America, 106,* 2074.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20(3),* 255-272.

Calandruccio, L., & Smiljanic, R. (2012). New Sentence Recognition Materials Developed Using a Basic Non-Native English Lexicon. *Journal of Speech, Language and Hearing Research, 55(5),* 1342.

Carney, D., Olson, K., Banaji, M., & Mendes, W. (2006). The faces of race-bias: Awareness of racial cues moderates the relation between bias and in-group facial mimicry. Unpublished manuscript, Harvard University, Cambridge, MA.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24(4),* 349.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12(2),* 163-170.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology, 36(3),* 316-328.

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. *Journal of Experimental Psychology: General, 134(2),* 222.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42(04),* 476-490.

Devos, T., & Banaji, M. R. (2005). American = white? *Journal of Personality and Social Psychology, 88(3),* 447.

Dodd, B., Plant, G., & Gregory, M. (1989). Teaching lip-reading: The efficacy of lessons on video. *British Journal of Audiology, 23(3),* 229-238.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82(1),* 62.

Drager, K. (2010). Sociophonetic variation in speech perception. *Language and Linguistics Compass, 4(7),* 473-480.

Edwards, A. L. (1957). The social desirability variable in personality assessment and research. NY: Dryden Press.

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders, 40(4),* 481.

Erber, N. P., & McMahan, D. A. (1976). Effects of sentence context on recognition of words through lipreading by deaf children. *Journal of Speech, Language and Hearing Research, 19(1),* 112.

Ferguson, S. H. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *The Journal of the Acoustical Society of America, 116,* 2365.

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language and Hearing Research, 50(5),* 1241.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20,* 303-315.

Fixmer, E., & Hawkins, S. (1998). The influence of quality of information on the McGurk effect. *Proceedings of AVSP'98,* 27-32.

Floccia, C., Butler, J., Goslin, J., & Ellis, L. (2009). Regional and foreign accent processing in English: Can listeners adapt? *Journal of Psycholinguistic Research, 38(4),* 379-412.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance, 32(5),* 1276.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America, 115,* 2246.

Gawronski, B. (2002). What does the Implicit Association Test measure? A test of the convergent and discriminant validity of prejudice-related IATs. *Experimental Psychology, 49(3),* 171-180.

Girin, L., Schwartz, J.-L., & Feng, G. (2001). Audio-visual enhancement of speech in noise. *The Journal of the Acoustical Society of America, 109,* 3007.

Goh, J. O., Suzuki, A., & Park, D. C. (2010). Reduced neural selectivity increases fMRI adaptation with age during face discrimination. *Neuroimage, 51(1),* 336-344.

Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108,* 1197.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102(1),* 4.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74(6),* 1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85(2),* 197.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97(1),* 17.

Harry, B., Davis, C., & Kim, J. (2012). Subliminal access to abstract face representations does not rely on attention. *Consciousness and Cognition, 21(1),* 573-583.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics, 48(4),* 865-892.

Hay, J., Nolan, A., & Drager, K. (2006a). From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review, 23(3),* 351-379.

Hay, J., Warren, P., & Drager, K. (2006b). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics, 34(4),* 458-484.

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication, 52(11),* 996-1009.

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America, 116,* 3108.

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America, 119,* 1740.

Heider, F. K., & Heider, G. M. (1940). A comparison of sentence structure of deaf and hearing children. *Psychological Monographs, 52(1),* 42-103.

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language and Hearing Research, 40(2),* 432.

Jokisch, O., Koloska, U., Hirschfeld, D., & Hoffmann, R. (2005). Pronunciation learning and foreign accent reduction by an audiovisual feedback system Affective Computing and Intelligent Interaction (pp. 419-425): Springer.

Junqua, J.-C., Fincke, S., & Field, K. (1999). The Lombard effect: A reflex to better communicate with others in noise. *Acoustic, Speech, and Signal Processing, 1999 (ICASSP'99) Proceedings, 4,* 2083–2086.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81(5),* 774.

Kidd Jr, G., Mason, C. R., Deliwala, P. S., Woods, W. S., & Colburn, H. S. (1994). Reducing informational masking by sound segregation. *The Journal of the Acoustical Society of America, 95,* 3475.

Kidd Jr, G., Mason, C. R., & Gallun, F. J. (2005). Combining energetic and informational masking for speech identification. *The Journal of the Acoustical Society of America, 118,* 982.

Kim, J., Sironic, A., & Davis, C. (2011). Hearing speech in noise: Seeing a loud talker is better. *Perception-London, 40(7),* 853.

Kinoshita, S., & Peek-O'Leary, M. (2005). Does the compatibility effect in the race Implicit Association Test reflect familiarity or affect? *Psychonomic Bulletin & Review, 12(3),* 442-452.

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication, 52(11),* 864-886.

Lombard, E. (1911). Le signe de l'elevation de la voix. *Annales Des Maladies De L'oreille, Du Larynx, Du Nez Et Du Pharynx, 37(101-119),* 25.

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage, 21(2),* 725-732.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 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.

Maison, D., Greenwald, A. G., & Bruin, R. H. (2004). Predictive validity of the Implicit Association Test in studies of brands, consumer attitudes, and behavior. *Journal of Consumer Psychology, 14(4),* 405-415.

Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle: The MIT Press.

Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(2),* 198-213.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27(7-8),* 953-978.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37(5),* 435-442.

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

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech, 49(1),* 101-112.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers, 36(4),* 630-633.

Montgomery, A. A., & Jackson, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *The Journal of the Acoustical Society of America, 73,* 2134.

Montgomery, A. A., Walden, B. E., & Prosek, R. A. (1987). Effects of consonantal context on vowel lipreading. *Journal of Speech, Language and Hearing Research, 30(1),* 50.

Munro, M. J. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition, 20(2),* 139-154.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45(1),* 73-97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38(3),* 289-306.

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

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18(1),* 62-85.

Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. *Social Cognition, 19(2)*, 97-144.

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician enhancement for speech-in-noise. *Ear and Hearing, 30(6),* 653-661.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech, Language and Hearing Research, 28(1),* 96.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language and Hearing Research, 29(4),* 434.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech, Language and Hearing Research, 32(3),* 600.

Pollack, I. (1975). Auditory informational masking. *The Journal of the Acoustical Society of America, 57,* S5.

Quillian, L. (2008). Does unconscious racism exist?. *Social Psychology Quarterly, 71(1),* 6-11.

Riley, R. (2008). Audio editing with Adobe audition: PC Publishing.

Roefs, A., & Jansen, A. (2002). Implicit and explicit attitudes toward high-fat foods in obesity. *Journal of Abnormal Psychology, 111(3),* 517.

Rogers, C. L., DeMasi, T. M., & Krause, J. C. (2010). Conversational and clear speech intelligibility of /bVd/ syllables produced by native and non-native English speakers. *The Journal of the Acoustical Society of America, 128,* 410.

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., & Abrams, H. B. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics, 27(3),* 465.

Rönnberg, J., Samuelsson, S., & Lyxell, B. (1998). Conceptual constraints in sentence-based lipreading in the hearing-impaired. *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory–visual Speech, 143-153.*

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22(2),* 318.

Rosini, L.-G. (1997). English with an accent: Language, ideology, and discrimination in the United States: Routledge.

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 environments. *Cerebral Cortex, 17(5),* 1147-1153.

Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the implicit association test. *Group Processes & Intergroup Relations, 10(3),* 359-372.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). " Unlearning" automatic biases: the malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology, 81(5),* 856.

Schwartz, J.-L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. The Journal of the Acoustical Society of *America, 127,* 1584.

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition, 93(2),* B69-B78.

Seferoğlu, G. (2005). Improving students' pronunciation through accent reduction software. *British Journal of Educational Technology, 36(2),* 303-316.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics. 21,* 427-444.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance, 28(6),* 1447.

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *The Journal of the Acoustical Society of America, 118,* 1677.

Smiljanić, R., & Bradlow, A. R. (2011). Bidirectional clear speech perception benefit for native and high-proficiency non-native talkers and listeners: Intelligibility and accentedness. *The Journal of the Acoustical Society of America, 130(6),* 4020.

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 and Hearing, 26(3),* 263-275.

Steele, C. M. (1997). A threat in the air: how stereotypes shape intellectual identity and performance. *American Psychologist, 52(6)*, 613.

Stein, B. E., Stanford, T. R., Ramachandran, R., Perrault Jr, T. J., & Rowland, B. A. (2009). Challenges in quantifying multisensory integration: Alternative criteria, models, and inverse effectiveness. *Experimental Brain Research, 198(2-3),* 113-126.

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage, 44(3),* 1210-1223.

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology, 18(1),* 86-100.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*, 212.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 335(1273),* 71-78.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech, 53(4),* 510-540.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102(4),* 1181-1186.

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research, 1111(1),* 134-142.

Weynand, D. (2010). Final Cut Pro 7: Pearson Deutschland GmbH.

Yi, H., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (in press). Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America.*