Reduced efficiency of audiovisual integration for nonnative speech

Han-Gyol Yi and Jasmine E. B. Phelps
Department of Communication Sciences & Disorders, The University of Texas at Austin, 2504A Whitis Avenue (A1100), Austin, Texas 78712
gyol@utexas.edu; jas.speechpath@gmail.com

Rajka Smiljanic
Department of Linguistics, The University of Texas at Austin, 305 East 23rd Street, Stop (B5100), Austin, Texas, 78712
rajka@austin.utexas.edu

Bharath Chandrasekaran[a)]
Department of Communication Sciences & Disorders, The University of Texas at Austin, 2504A Whitis Avenue (A1100), Austin, Texas 78712
bchandra@austin.utexas.edu

Abstract: The role of visual cues in native listeners' perception of speech produced by nonnative speakers has not been extensively studied. Native perception of English sentences produced by native English and Korean speakers in audio-only and audiovisual conditions was examined. Korean speakers were rated as more accented in audiovisual than in the audio-only condition. Visual cues enhanced word intelligibility for native English speech but less so for Korean-accented speech. Reduced intelligibility of Korean-accented audiovisual speech was associated with implicit visual biases, suggesting that listener-related factors partially influence the efficiency of audiovisual integration for nonnative speech perception.

PACS numbers: 43.70.Mn, 43.71.Bp, 43.71.Gv [AC]
Date Received: July 19, 2013 **Date Accepted:** September 12, 2013

1. Introduction

Visual cues aid speech perception for clear and intact auditory stimuli (Lander and Capek, 2013). Temporal and phonemic cues in visual speech also help segregate an auditory stream from competing background noise (Sumby and Pollack, 1954; Grant and Seitz, 2000). However, our knowledge of audiovisual (AV) integration is mostly limited to the perception of speech produced by native speakers (for work on perception of native AV speech by nonnative listeners, see Hardison, 2003; Wang *et al.*, 2009).

AV integration for nonnative speech may differ from native speech due to speaker- and listener-related factors (Hazan *et al.*, 2010). With respect to speaker-related factors, nonnative visual cues deviate from familiar articulatory patterns, reducing the effectiveness of this intelligibility-enhancing cue (unfamiliar regional accents; Irwin *et al.*, 2011). Regarding listener-related factors, visual cues convey information that could enhance the perceived nonnativeness of the speaker. For example, East Asian faces are less likely to be assumed to be native to the American English–speaking environment (Devos and Banaji, 2005). This assumption may alter the listener's perception by exaggerating the perceived nonnativeness of an East Asian speaker (Drager, 2010). Indeed, native English listeners incorporate visual cues more

[a)]Author to whom correspondence should be addressed. Also at: Institute for Neuroscience, The University of Texas at Austin, Austin, TX and Center for Perceptual Systems, The University of Texas at Austin, Austin, TX.

when they are listening to CV syllables produced by native Mandarin Chinese speakers relative to native English speakers (Hazan *et al.*, 2010). However, it is unclear how this increased visual weighting for nonnative speakers affects sentence intelligibility in challenging listening environments.

To this end, we examined the effect of visual cues on nonnative speech perception in two domains: speech intelligibility in noise and accent ratings. Native American English listeners participated in a speech perception in noise task for sentences produced by native English and native Korean speakers, with and without visual cues. Visual enhancement for speech intelligibility was predicted for both speaker groups. However, the nonnative visual benefit was expected to be less. This was due to the deviance from native visual speech patterns (speaker-related effect) and the exaggerated perception of nonnativeness due to the East Asian facial cues (listener-related effect). An implicit association test (IAT) (Devos and Banaji, 2005) was administered to examine the listener-related effect. We expected the IAT scores to predict the extent of relative difficulty for nonnative speech perception in noise when visual cues are available. A separate group of participants provided accentedness ratings for the same set of stimuli. This was done to ascertain that the nonnative speech stimuli were perceived to be accented, and to examine the effect of visual cues on subjective rating of accentedness.

2. Methods

2.1 Participants

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

2.2 Materials

AV stimuli. Forty target sentences with four keywords (e.g., "The GIRL LOVED the SWEET COFFEE"; Calandruccio and Smiljanic, 2012) were produced by two native American English (one female) and two native Korean speakers (one female). These target sentences were mixed with random samples of a six-talker babble track created from 30 simple, meaningful sentences (Bradlow and Alexander, 2007) produced by six native speakers of American English [three female; Van Engen *et al.*, 2010; supplementary materials; Fig. 1(a)]. The noise tracks, accompanied by a freeze frame of the

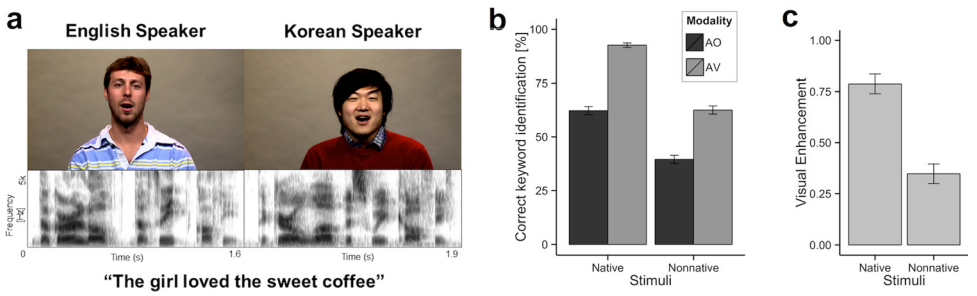

Fig. 1. (Color online) (a) Visual (upper panel) and auditory (lower panel) speech cues of the sentence "the girl loved the sweet coffee" produced by native and non-native speakers. The sample AV stimuli are available as supplementary materials (Mm. 1; Mm. 2). (b) Percentage of the keywords correctly identified for the speech perception in noise task for English (left bars) and Korean (right bars) speakers, without (darker fill) and with (lighter fill) visual cues. (c) Visual enhancement measures [(AV−AO)/(1−AO)] compared between native English and Korean speakers.

video, enveloped the target sentences by 500 ms. The signal-to-noise ratio was set to −4 dB (target RMS amplitude = 65 dB) based on pilot studies to avoid ceiling or floor performance. The same target sentences without noise were used for accent ratings.

Mm. 1. Sample AV stimulus produced by a native English speaker. This file has been down-sampled from the original format (video: 29.97 fps, 1920 × 1080, DV-PAL, 4.2 MB/s; audio: 22 050 Hz, 16-bit). This is a file of type "avi" (3.5 MB).

Mm. 2. Sample AV stimulus produced by a native Korean speaker. This is a file of type "avi" (2.7 MB).

IAT stimuli. Ten young adult Asian (five female) and ten Caucasian (five female) face images were used for Caucasian vs Asian face categories (Minear and Park, 2004). Public domain images of ten iconic American scenes (e.g., Grand Canyon, Statue of Liberty) and ten non-American foreign scenes (e.g., Eiffel Tower, Angkor Wat) were obtained online and used for American vs foreign scene categories.

2.3 Procedures

Speech perception in noise. Participants were placed in a sound-attenuated booth and presented with 40 speech-in-noise stimuli. Each sentence was randomly produced once by only one of four speakers, in either audio-only (AO) or AV condition. In the AO condition, the video track was replaced with a fixation cross. The presentation order was randomized. After each stimulus presentation, participants typed the target sentences using a computer keyboard, which were scored for by-keyword accuracy.

Accent rating. Participants were placed in a sound-attenuated booth and listened to all 40 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. Using a computer keyboard, participants rated how accented each sentence was on a 1-to-9 Likert scale: 1 = no foreign accent; 9 = very strong foreign accent (Smiljanic and Bradlow, 2011).

IAT. Participants were instructed to perform a response time task in which they were to respond as quickly as possible without sacrificing accuracy. 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. An incorrect response led to a corrective feedback of "Error!" [Fig. 2(a)].

2.4 Data analysis

Speech in Noise Perception. SPIN outcome (correct vs incorrect) for each word response was entered as the dependent variable using a binomial logit link (Bates *et al.*, 2012). Fixed effects included modality and nativeness, with by-subject, by-sentence, and by-word random intercepts.

Visual benefit and native-speaker benefit were also measured by using an established method of calculating enhancement (Sommers *et al.*, 2005): AV boost = (AV − AO)/(1 − AO); native boost = (native − nonnative)/(1 − nonnative).

Accent ratings. Accent rating scores were converted to continuous percentage scale of native-like accent: 0%: least native-like; 100%: most native-like. A linear mixed effects analysis (Bates *et al.*, 2012) was run on the percentages as the dependent variable. Fixed effects were modality condition (AV vs AO) and nativeness of the speaker (English vs Korean), with by-subject and by-sentence random intercepts. *p*-values of the fixed effects were calculated with Markov Chain Monte Carlo sampling ($n = 10\,000$).

IAT. Response times (RT) were scored to yield one IAT score per participant. A higher IAT score indicated a greater implicit bias towards making Caucasian-

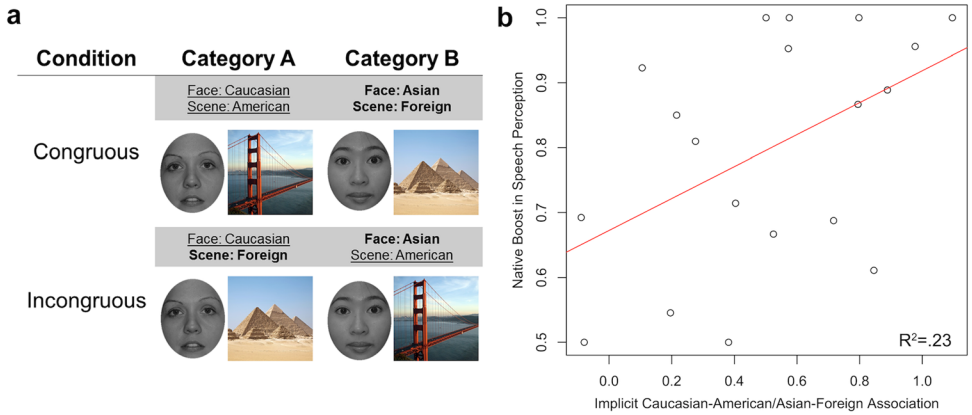

Fig. 2. (Color online) Implicit association test. (a) Face (ten Caucasian; ten Asian) and scene (ten American; ten foreign) images were presented. In the congruous condition, participants were instructed to group Caucasian faces and American scenes together, and Asian faces and foreign scenes together. 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 are available positively correlate with each other, $r(17) = 0.482$, $p = 0.037$, $R^2 = 0.23$.

American and Asian-foreign associations (Devos and Banaji, 2005). An outlier analysis was performed ($< \pm$ 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 ($n = 10\,000$).

3. Results

3.1 Speech perception in noise

Percentages of the keywords correctly identified for English speakers were 62.4% in AO and 92.9% in AV. Percentages of the keywords correctly identified for Korean speakers were 39.5% in AO and 62.5% in AV [Fig. 1(b)]. Comparison of AV/AO ratios based on raw values suggests a 48% increase for English and 58% for Korean speakers. However, this method is biased against the native AO condition, which yielded higher scores than the nonnative AO condition. The percentage values represent the average probability that each word in a sentence will be perceived correctly in a given condition. The null distribution in this case follows the binomial distribution of "correct" or "incorrect." As performance reaches the positive extreme, the null probability associated with performance exponentially decreases. Therefore, a linear comparison of simple ratios of percentage scores is inadequate. The analytic method must take account of 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 (Sommers *et al.*, 2005; Grant and Seitz, 2000). The visual enhancement for native speech ($M = 0.79$; $SD = 0.18$) was higher than for nonnative speech ($M = 0.35$; $SD = 0.32$), $t(20) = 6.49$, $p < 0.0001$ [Fig. 1(c)], indicating that the visual cues benefit native speech more than nonnative speech.

A more direct approach is 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 last interaction term is of main interest in this paper. 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 $= 0.2621$, $z = 2.65$, $p = 0.008$. The nativeness effect was significant, $b = -1.1925$, SE $= 0.1134$, $z = -10.51$, $p < 0.0001$, 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 $= 0.1624$, $z = 13.40$, $p < 0.0001$, such that keywords were more correctly identified in AV than in AO. The nativeness by condition interaction effect was significant, $b = -1.1088$, SE $= 0.1974$, $z = -5.62$, $p < 0.0001$, 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 AV integration for perception of nonnative speech relative to native speech.

3.2 Accent rating

The average native-like rating for English speakers was 96.2% in the AO condition and 97.1% in the AV condition. The average native-like rating for Korean speakers was 20.7% in AO and 18.9% in AV. The intercept was significant, $b = 96.1725$, SE $= 2.0581$, $t = 46.73$, $p < 0.0001$. The nativeness effect was significant, $b = -75.4605$, SE $= 0.8258$, $t = -91.38$, $p < 0.0001$, with Korean speakers rated as more foreign-accented. The modality condition effect was not significant, $b = 0.9324$, SE $= 0.8258$, $t = 1.13$, $p = 0.2630$. The nativeness by modality condition interaction effect was significant, $b = -2.7173$, SE $= 1.1675$, $t = -2.33$, $p = 0.0166$, that Korean speakers were rated as more foreign-accented in AV relative to AO, while English speakers were rated to be less accented in AV than in AO.

3.3 IAT

On average, IAT scores from native English listeners were higher than zero ($M = 0.511$; SD $= 0.347$), $t(18) = 6.41$, $p < 0.0001$, indicating participants had an overall bias toward congruous associations. Individual IAT scores were positively correlated with native boost in AV, $r(17) = 0.482$, $p = 0.037$, 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 [Fig. 2(a)]. In contrast, IAT scores were not significantly correlated with native boost in AO, $r(17) = 0.064$, $p = 0.80$, indicating that the bias against incongruous associations was not related to relative performance across sentences produced by English and Korean speakers in AO.

Next, the linear mixed effects analyses allowed us 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. The intercept was significant, $b = 858.33$, SE $= 41.77$, $t = 20.55$, $p < 0.0001$. The incongruous condition showed a significant effect, $b = 174.11$, SE $= 15.86$, $t = 10.97$, $p < 0.0001$, indicating that the responses in the incongruous condition were slower than in the congruous condition by approximately 174 ms. Second, the model with category condition and SPIN native boost scores (AV) was run. The intercept was significant, $b = 1059.77$, SE $= 200.94$, $t = 5.27$, $p < 0.0001$. The incongruous condition effect was not significant, $b = -44.07$, SE $= 76.66$, $t = -0.58$, $p = 0.58$, nor was the SPIN native boost effect, $b = -254.92$, SE $= 248.81$, $t = -1.03$, $p = 0.31$. However, there was a significant interaction between the incongruous condition and the SPIN native boost scores, $b = 276.10$, SE $= 94.92$, $t = 2.91$, $p = 0.0024$. 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.

4. Discussion

We examined the role of visual cues in processing native and nonnative speech in adverse listening conditions. Visual cues significantly enhanced speech intelligibility in noise for both native and nonnative speech. However, visual enhancement was greater for native speakers than for nonnative speakers. Both the speaker-related effect of signal degradation in visual speech cues (Irwin *et al.*, 2011) as well as the listener-related effect of implicit association with nonnativeness (Devos and Banaji, 2005) may have contributed to the reduced efficiency of AV integration for nonnative speech.

Two results from this study suggest that listener-related factors contribute to the reduced efficiency of AV integration in nonnative speech perception. First, visual cues led Korean speakers to be rated as more accented, and English speakers as less accented. Second, the participants with stronger Caucasian-American and Asian-foreign associations had greater relative difficulty with nonnative speech in the AV condition, but not in the AO condition. These results in conjunction provide a preliminary insight into how visual cues provide speaker information, which can interact with a listener's non-linguistic visual bias, affecting speech processing.

Visual cues provide facial information about the speaker that can be processed in the absence of conscious allocation of attention (Harry *et al.*, 2012). Non-Caucasian faces may provide a socioindexical cue regarding increased nonnativeness (Devos and Banaji, 2005). These facial cues could lead to altered incorporation of visual cues (Drager, 2010) since listeners are able to automatically adjust phonetic perception to indexical information (McQueen *et al.*, 2006). In the current study, facial cues resulted in reduced visual benefit for nonnative speakers. This interpretation does not imply that native listeners reduce their reliance on visual cues in nonnative speech perception (Hazan *et al.*, 2010), but rather that they are not able to adequately benefit from nonnative visual cues regardless of the degree of their reliance.

The current efforts to address difficulties arising from nonnative speech perception primarily focus on reduction or elimination of nonnative speakers' accents. The listeners' role in resolving nonnative speech patterns are not addressed (Derwing and Munro, 2009). Given that visual cues are typically available in everyday speech communication, it is valuable to consider their effects on nonnative speech perception. The results from this study confirm that native listeners can incorporate visual cues to improve nonnative speech intelligibility. However, the findings from this study indicate that the visual benefit is less for nonnative than for native speech. While this inefficient AV nonnative speech processing is believed to be caused in part by the degraded nature of the nonnative speech signals, the correlational analysis provides evidence that listener-related factors also play a significant role.

Acknowledgments

This research was funded by Longhorn Innovation Fund for Technology, awarded to B.C. and R.S. The authors thank Kristin J. Van Engen for providing invaluable assistance in stimulus preparation and data analysis insight. The authors also thank the SoundBrain Lab research assistants for data collection: Blue Alozie, Tiffany Berry, Kadee Bludau, Millicent Campbell, Kathryn Curry, Dionne Dias, Morgan Elkins, Sarah Evans, Kim Kowinski, Carolyn Linebaugh, and Elsa Tran.

References and links

Bates, D., Maechler, M., and Bolker, B. (**2012**). "lme4: Linear mixed-effects models using S4 classes" [computer program].

Bradlow, A. R., and Alexander, J. A. (**2007**). "Semantic and phonetic enhancements for speech-in-noise recognition by native and nonnative listeners," J. Acoust. Soc. Am. **121**(4), 2339–2349.
Calandruccio, L., and Smiljanic, R. (**2012**). "New sentence recognition materials developed using a basic nonnative English lexicon," J. Speech. Lang. Hear. R. **55**(5), 1342–1355.
Derwing, T. M., and Munro, M. J. (**2009**). "Putting accent in its place: Rethinking obstacles to communication," Lang. Teach. **42**(04), 476–490.
Devos, T., and Banaji, M. R. (**2005**). "American = White?," J. Pers. Soc. Psychol. **88**(3), 447–466.
Drager, K. (**2010**). "Sociophonetic variation in speech perception," Lang. Linguist. Compass. **4**(7), 473–480.
Grant, K. W., and Seitz, P. F. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences," J. Acoust. Soc. Am. **108**(3), 1197–1208.
Hardison, D. (**2003**). "Acquisition of second-language speech: Effects of visual cues, context, and talker variability," Appl. Psycholinguist. **24**(4), 495–522.
Harry, B., Davis, C., and Kim, J. (**2012**). "Subliminal access to abstract face representations does not rely on attention," Conscious. Cogn. **21**(1), 573–583.
Hazan, V., Kim, J., and Chen, Y. (**2010**). "Audiovisual perception in adverse conditions: Language, speaker and listener effects," Speech Commun. **52**(11), 996–1009.
Irwin, A., Pilling, M., and Thomas, S. M. (**2011**). "An analysis of British regional accent and contextual cue effects on speechreading performance," Speech Commun. **53**(6), 807–817.
Lander, K., and Capek, C. (**2013**). "Investigating the impact of lip visibility and talking style on speechreading performance," Speech Commun. **55**(5), 600–605.
McQueen, J. M., Norris, D., and Cutler, A. (**2006**). "The dynamic nature of speech perception," Lang. Speech. **49**(1), 101–112.
Minear, M., and Park, D. C. (**2004**). "A lifespan database of adult facial stimuli," Behav. Res. Methods Instrum. Comput. **36**(4), 630–633.
Smiljanic, R., and Bradlow, A. R. (**2011**). "Bidirectional clear speech perception benefit for native and high-proficiency nonnative talkers and listeners: Intelligibility and accentedness," J. Acoust. Soc. Am. **130**(6), 4020–4031.
Sommers, M. S., Tye-Murray, N., and Spehar, B. (**2005**). "Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults," Ear Hear. **26**, 263–275.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**(2), 212–215.
Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., and Bradlow, A. R. (**2010**). "The Wildcat Corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles," Lang. Speech.**53**, 510–540.
Wang, Y., Behne, D. M., and Jiang, H. (**2009**). "Influence of native language phonetic system on audio-visual speech perception," J. Phon. **37**(3), 344–356.