PERCEPTUAL ADAPTATION TO STEREOTYPED ACCENTS IN AUDIO-VISUAL SPEECH

Molly Babel and Gloria Mellesmoen

Department of Linguistics, University of British Columbia
molly.babel@ubc.ca, gloria.mellesmoen@alumni.ubc.ca

ABSTRACT

Listeners entertain hypotheses about how social characteristics affect a speaker's pronunciation. While some of these hypotheses may prove truthful, thus facilitating spoken language processing, others may be erroneous stereotypes that impede comprehension. For example, there are a range of studies which show that listeners' stereotypes of language and ethnicity pairings in varieties of North American English can improve intelligibility and comprehension or hinder these processes. Using audio-visual speech we examine how listeners adapt to speech in noise from four speakers who are representative of the accent-ethnicity stereotypes in the local speech community: an Asian native English speaker, a Caucasian native English speaker, an Asian non-native English speaker, and a Caucasian non-native English speaker. The results suggest that accent-ethnicity pairings that are incongruent with local stereotypes may inhibit adaptation to speech in noise.

Keywords: Perceptual adaptation, social stereotypes, speech in noise, sociophonetics

1. INTRODUCTION

Listeners' experiences in the linguistic world contribute to the formation and reinforcement of associations between language and society. Listeners learn that, for example, females, on average, have smaller vocal tracts than men, and thus generally have higher frequency boundaries between vowels [11] and sibilant fricatives [17]. Such expectations about the relationship between talker size and phonetic realizations arguably assist in processing spoken language more efficiently and adeptly. While associations related to gender or sex are partially rooted in physiological differences (as opposed to culturally-specific learned patterns, see [10]) between women and men, listeners also connect pronunciation patterns with more arbitrary social groups. Drawing upon learned associations, listeners can categorize a speaker by a number of different social identities (e.g., gender, ethnicity, social class, etc., see [4, 7]).

1.1. Stereotypes and Expectations

These expectations and the sociolinguistic knowledge listeners carry can warp their perception of the speech stream. Listeners' ultimate percepts or decisions about what they heard of a given utterance are influenced by what they *expect* a talker from a particular social category to produce. For example, given acoustically identical perceptual stimuli, New Zealand listeners perceive speakers who seem younger as having a more complete NEAR/SQUARE merger, consistent with them being probabilistically more likely to have merged the sounds [8]. Niedzielski [15] found that listeners from Michigan, USA assumed that an apparent speaker from Ontario, Canada had a different accent from their own (despite this lack of difference) and categorized vowels accordingly. Niedzielski also showed that these Michigan listeners perceived their own accent as patterning more with a mainstream American one, indicating a disconnect between actual and perceived pronunciation in their speech community. This indicates that listener expectations about accents and speech patterns, including their own, affect their perceptual space.

Listener associations between accent and ethnicity in English-speaking North America present a particular challenge, as the associations are frequently shown to be fallible. Despite multicultural and diverse non-white demographics, to be considered maximally "American", one must be Caucasian [3]. This association is implicated in speech studies by who is expected to speak "unaccented" English. For example, an influential set of studies paired photos of a Caucasian face and a East Asian face with voices representing native and non-native accents [16, 12]. When the voices were paired with the East Asian face, they were perceived as more accented and were associated with lower accuracy on a cloze task. Kang and Rubin reason that *reverse linguistic stereotyping* results in evaluations of low social status negatively affecting speech comprehension.

Independent of social prestige, experience and stereotypes may affect speech processing. McGowan found that Mandarin-accented English was

more intelligible when paired with an East Asian face than with a Caucasian face [14]. This facilitating effect in comprehending L2-accented speech is consistent with expectations about the phonetic patterns associated with a given social group. Using speech from a larger set native speakers of Canadian English, Babel and Russell demonstrated a similar effect in a speech in noise task that compared audio-only trials with ones pairing audio with Caucasian-Canadian or Chinese-Canadian faces [1]. They found lower accuracy in the transcription of Chinese-Canadian speech only in combination with Chinese-Canadian faces. This effect was greater for listeners who reported spending more time with Chinese Canadians, suggesting that the findings may not be about negative social associations, but instead involve erroneous ethnicity/language expectations.

Generally, audio-visual speech receives a boost in performance compared to audio-only speech (eg., [19]). This audio-visual benefit, however, has been shown to be larger for natively accented talkers [20]. Yi and colleagues tested listeners using native and Korean-accented English in audio-only and audio-visual conditions. A greater audio-visual boost was found for the native English speakers. For the Korean-accented speakers, listeners' performance was predicted by the strength of an association between the categories "Asian" and "foreign". They conclude less experience with Korean faces inhibits listener ability to exploit the facial movements that are known to aid alignment and boost intelligibility.

1.2. Hypotheses and Predictions

The results described above illustrate that listeners use experiences and stereotypes to buffer expectations that help and hinder the processing of novel voices. The literature suggests that listeners should be better at adapting to accent-ethnicity associations that match local stereotypes. We test this with a speech in noise sentence transcription task using naturally produced audio-visual stimuli with speech embedded in -5 dB SNR pink noise from four talkers who vary in terms of self-identified ethnicity – Caucasian and Asian – and whether they speak English as a first or second language. Comparing high predictability training sentences and low predictability test sentences, we expect listeners to adapt more easily to the talkers who match accent-ethnicity stereotypes. Thus, we expect to see a reduction in transcription accuracy between training and test trials for the Asian native and the Caucasian non-native English speakers due to assumptions that ethnically Asian individuals should be non-native English speakers and ethnically Caucasian individuals should be native speakers of English. Being trained on an accent-ethnicity pairing counter to local stereotypes is predicted to make adaptation more difficult due to a mismatch between predicted and perceived signals. The Caucasian native English and the Asian non-native English talkers conform to local stereotypes, and we predict that listeners will adapt more to these talkers, showing some improvement in transcription accuracy between the high predictability training sentences and the low predictability test set.

2. METHODOLOGY

2.1. Materials

2.1.1. Audio-Visual Stimuli

Four female talkers in their twenties were recorded reading high and low predictability sentences from [2]. The talkers represented local accent-ethnicity stereotypes and included two native speakers of Canadian English and two non-native English speakers. For both the native and non-native pairs, one talker was Asian and the other was Caucasian. The Asian non-native English talker was a native speaker of Mandarin and the Caucasian one was a native speaker of Spanish; these speakers were chosen out of convenience.

Audio recordings were digitized at 44.1 kHz using a Sennheiser MKH-416 shotgun microphone connected to a USB Pre-2 amplifier and a PC. Video recordings were made using Panasonic HC-V700M high definition video camera, which also recorded audio. The video recordings included the talkers from the neck up against a white background. The high quality audio recordings were RMS amplitude normalized and embedded in pink noise at a -5 dB SNR. The video and high-quality audio streams were synced using Adobe Premier Pro using the lower quality audio recorded from the video recorder. Sentences with speech errors were eliminated, leaving 120 unique sentences. In an audio-only procedure, listeners heard the sentences while looking at a blank screen. In an audio-visual condition, participants heard audio while watching the accompanying video. In this paper, we set aside the audio-only data and focus on the audio-visual condition in order to test our hypotheses about accent and ethnicity stereotypes.

2.2. Participants

A total of 83 listeners were recruited from undergraduate linguistics courses and received partial

course credit in exchange for their participation. There were 66 female and 17 male participants between 18 and 26 years of age (Mean = 20). Listeners were either native or early learners of English, which we operationalize as before the age of 5.

2.3. Procedure

Participants were seated in front of a computer in sound attenuated cubicles for the duration of the experiment. Listeners heard each sentence over headphones at approximately 65 dB while watching accompanying video of the talker on the screen. They were asked to type sentences on a keyboard and told to focus on being as accurate as possible while not worrying about minor spelling errors.

The task was blocked by talker, and listeners heard 30 sentences from each talker. In order to control for talker order, there were 24 different permutations of the experiment. These orders were implemented cyclically, such that one participant would have order A and the next order B, resulting in approximately three to four participants for each. The 30 sentences were separated into 15 high predictability and 15 low predictability blocks, randomly selected for each listener. The high and low predictability blocks were thus designed and then analyzed as training and test blocks, respectively. There were breaks between talkers, but not between sentence types within a talker.

This within-subject design for all talkers allows us to ignore talker-specific differences in intelligibility and focus on change – improvement or decline in performance – between high and low predictability blocks for each of the four talkers.

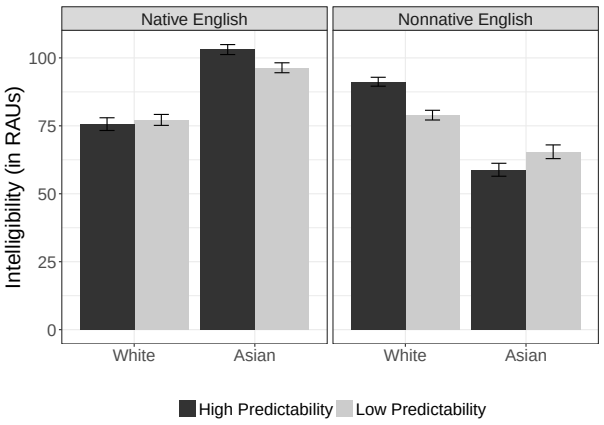
Participants also completed a modified version of LexTale [13] and an Implicit Association Test [5] designed to assess the association of Canada and Caucasian faces compared to Foreign entities and Asian faces, similar to [20]. Due to time and space constraints, these data have not been analyzed and will not be discussed further in this paper.

3. ANALYSIS AND RESULTS

The measure of interest in this study is the change in listeners' accuracy in transcribing a talker's speech in noise between the set of high predictability sentences and the set of low predictability sentences. Transcription accuracy was assessed as the number of correct words per target sentence. Transcription spelling was first automatically corrected using spell-check and then hand-corrected. A Python script assigned one point to each correct word and incorrect words were not penalized. All contractions were hand-checked, treated as two words, and scored accordingly. For example, the contraction "it's" in the sentence "dad thinks that it's funny" would be scored as if it were "it is", meaning that the sentence would receive a score of six for the target sentence "dad thinks that it is funny". Otherwise, for a correct score, the transcribed word needed to match the target exactly. In order to avoid assumptions about listener intentions, words with the wrong tense or number inflection were treated as incorrect.

Transcription accuracy was normalized to Rationalized Arcsine Units (RAUs) following [18] and used as the dependent measure in a linear mixed effects model. Predictability (High, Low with High as the reference level), Talker Native Language (Native English, Nonnative English with Native English as the reference level), Talker Ethnicity (Caucasian, Asian with Caucasian as the reference level) were entered as fixed effects with all possible interactions. Subject and Sentence were random effects with Predictability, Talker Native Language, and Talker Ethnicity as random slopes for the Subject intercept.

Figure 1: Intelligibility, shown in normalized RAU values, for the four talkers separated by high and low predictability sentences.



Effects with t-values of greater than |2| are interpreted here as significant. The model intercept returned as as significant [$\beta = 76.8, SE = 3.17, t = 24.3$]. There were simple effects of Talker Native Language [$\beta = 14.42, SE = 1.91, t = 7.54$] and Talker Ethnicity [$\beta = 25.9, SE = 1.97, t = 13.17$]. Given that the reference levels were high predictability, Native English, and Caucasian, these effects indicate that the Nonnative speakers and Asian speakers were more intelligible than the Caucasian Native English speaker. These simple effects were overshadowed by two-way interactions of Predictability and Talker Native Language [$\beta =$

$-12.21, SE = 2.29, t = -5.33$], Predictability and Talker Ethnicity [$\beta = -6.04, SE = 2.11, t = -2.87$], and Talker Native Language and Talker Ethnicity [$\beta = -59.05, SE = 2.79, t = -21.14$]. The three-way interaction between Predictability, Talker Native Language, and Talker Ethnicity was also significant [$\beta = 25.33, SE = 3.33, t = 7.6$]. Group means with by-subject standard error for this three-way interaction are shown in Fig. 1. As can be seen in this figure, the four talkers varied in their overall intelligibility with the Caucasian Native English speaker having surprisingly low intelligibility.

Our design allows for these differences in baseline intelligibility for the talkers by having the high and low predictability sentences as training and test blocks for each listener. Thus, to more straightforwardly test within-talker changes in performance between the high and low predictability blocks, separate analyses were run for each talker with Predictability as a fixed effect and Subject and Sentence as random effects. Predictability was included as a by-Subject random slope. These results largely confirm what is visible in Figure 1.

The White English L1 speaker showed nearly no change [$\beta = 0.32, SE = 4.7, t = 0.07$] between the high ($M = 76$ RAUs, $SD = 47$) and low ($M = 77$ RAUs, $SD = 40$) predictability sentences. The more intelligible Asian English L1 speaker showed significant reduction [$\beta = -7.25, SE = 3.53, t = -2.06$] in intelligibility in the low predictability sentence set ($M = 103$ RAUs, $SD = 34$) compared to the high predictability set ($M = 96$ RAUs, $SD = 35$). The White English L2 speaker showed a predicted decline [$\beta = -12.34, SE = 4.26, t = -2.9$] in intelligibility between the high ($M = 91$ RAUs, $SD = 39$) and low ($M = 79$ RAUs, $SD = 38$) sentence sets. The Asian English L2 speaker showed non-significant improvement between the high ($M = 59$ RAUs, $SD = 50$) and low ($M = 65$ RAUs, $SD = 47$) predictability sentence blocks [$\beta = 6.59, SE = 5.25, t = 1.3$].

4. DISCUSSION

Listeners' ability to transcribe speech in noise declined significantly for the Asian Native and Caucasian Non-native English speakers between the high predictability training sentences and the low predictability test sentences. Transcription accuracy trended towards improvement for the Asian Non-native speaker and barely changed for the Caucasian Native English speaker. The results suggest that stereotypes about accent and ethnicity play a role in perceptual adaption, particularly for speakers who do not match local stereotypes of who is expected to be a native speaker of English. The discrepancy between socially-based expectations and reality appears to have hindered listener adaptation, resulting in reduced intelligibility for the Asian native English speaker and the Caucasian nonnative English speaker. Though we predicted some positive adaption for the talkers who matched accent-ethnicity stereotypes, there was no significant improvement for either the Asian nonnative English or the Caucasian native English speaker. This suggests that while alignment between talker accent-ethnicity and social expectations does not interfere with adaptation, it does not necessarily facilitate it in adverse listening conditions (see also [6]).

Though the present study focuses on talker characteristics, listener attributes may also underlie performance. Possible mediating listener-based factors include linguistic flexibility, multilingual exposure, and rigidity of social expectations. We collected data, but have yet to analyze it, that pertain to these listener factors. In a study on older listeners, [9] found those with higher vocabularies showed better adaptation to novel accents. Our modified version of LexTale [13] will allow us to test how vocabulary size affects adaptation within an undergraduate-aged population. We also have the data to test listeners' associations between foreignness and Canadianness – specifically, we expect that listeners with a stronger Asian=Foreign bias (as reflected on an Implicit Association Test [5]) will be more susceptible to the interference of accent-ethnicity stereotypes. Flexibility in social expectations may promote increased intelligibility through reducing the potential interference of a mismatch between accent-ethnicity stereotypes and talker characteristics. The results reported here mask a large amount of variation between listeners, which we hope to account for with these measures.

5. CONCLUSION

Low predictability sentences produced by talkers with ethnicity-accent associations that do not match local stereotypes – Asian native English and Caucasian non-native English speakers – were transcribed less accurately following exposure to high predictability sentences. We hypothesize that this lack of adaptation is due to a mismatch of expectations which inhibits learning of speech in noise. These results contribute to the literature on accent-ethnicity stereotypes in the North American context [12, 16, 14, 1].

6. REFERENCES

[1] Babel, M., Russell, J. 2015. Expectations and speech intelligibility. *The Journal of the Acoustical Society of America* 137(5), 2823–2833.

[2] 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(4), 2339–2349.

[3] Devos, T., Banaji, M. R. 2005. American= white? *Journal of personality and social psychology* 88(3), 447.

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

[5] 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.

[6] Hanulíková, A. 2018. The effect of perceived ethnicity on spoken text comprehension under clear and adverse listening conditions. *Linguistics Vanguard* 4(1).

[7] Hay, J., Drager, K. 2007. Sociophonetics. *Annu. Rev. Anthropol.* 36, 89–103.

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

[9] Janse, E., Adank, P. 2012. Predicting foreign-accent adaptation in older adults. *The Quarterly Journal of Experimental Psychology* 65(8), 1563–1585.

[10] Johnson, K. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of phonetics* 34(4), 485–499.

[11] Johnson, K., Strand, E. A., D'Imperio, M. 1999. Auditory–visual integration of talker gender in vowel perception. *Journal of phonetics* 27(4), 359–384.

[12] Kang, O., Rubin, D. L. 2009. Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology* 28(4), 441–456.

[13] Lemhöfer, K., Broersma, M. 2012. Introducing lextale: A quick and valid lexical test for advanced learners of english. *Behavior research methods* 44(2), 325–343.

[14] McGowan, K. B. 2015. Social expectation improves speech perception in noise. *Language and Speech* 58(4), 502–521.

[15] Niedzielski, N. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of language and social psychology* 18(1), 62–85.

[16] Rubin, D. L. 1992. Nonlanguage factors affecting undergraduates' judgments of nonnative english-speaking teaching assistants. *Research in Higher education* 33(4), 511–531.

[17] Strand, E. A., Johnson, K. 1996. Gradient and visual speaker normalization in the perception of fricatives. *KONVENS* 14–26.

[18] Studebaker, G. A. 1985. A rationalized arcsine transform. *Journal of Speech, Language, and Hearing Research* 28(3), 455–462.

[19] Sumby, W. H., Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america* 26(2), 212–215.

[20] Yi, H.-G., Phelps, J. E., Smiljanic, R., Chandrasekaran, B. 2013. Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America* 134(5), EL387–EL393.