Speakers talk more clearly when they see an East Asian face: Effects of visual guise on speech production

Nicholas Aoki and Georgia Zellou

University of California, Davis
nbaoki@ucdavis.edu, gzellou@ucdavis.edu

ABSTRACT

Extensive work in speech perception indicates that given the same speech signal, listeners behave differently when viewing East Asian faces compared to Caucasian faces. An untested question is whether visual guise also affects *speech production*. If speakers assume that an Asian face depicts a non-native English speaker, we predict that speech towards an Asian face should be hyper-articulated compared to speech towards a Caucasian face and should acoustically resemble speech towards an imagined non-native listener. Moreover, individual differences in ratings of how likely the faces depict non-native speakers should predict variation. Results reveal that: (1) speech towards an East Asian face is hyper-articulated compared to a Caucasian face through a slower speaking rate; (2) non-native-directed speech is even more hyper-articulated than speech towards an East Asian face; (3) ratings do not predict differences between visual conditions. This study has implications for the relationship between speech perception, production, and social expectations.

Keywords: Speaking style, Visual guise, Non-native-directed speech, Perceived ethnicity, Sociolinguistics

1. INTRODUCTION

Extensive research has shown that the perceived ethnicity of visual guises can impact speech perception by triggering sociolinguistic expectations about speakers [1, 2, 3, 4]. For example, Rubin [5] presented listeners with the same recording of a lecture produced by a native speaker of American English and found that listeners who see a picture of an East Asian woman show lower comprehension (i.e., lower cloze test accuracy) of the lecture material and rate the voice as more accented compared to listeners who see a picture of a Caucasian woman. This phenomenon has been explained as a type of linguistic stereotyping – visual cues to a speaker's group membership can generate expectations about their speech as being more accented due to listener biases (e.g., that East Asians are non-native English speakers) [5, 6, 7, 8]. Other work has found that an East Asian guise can enhance comprehension. McGowan [9] showed that for Mandarin-accented speech, transcription accuracy in a speech-perception-in-noise task is higher when a picture of an East Asian woman is presented to listeners compared to a Caucasian woman or a silhouette. The greater congruency of the East Asian face with accented speech thus increases comprehension compared to the less compatible Caucasian face [9, 10]. In general, these prior studies in speech perception are consistent with the assumption that seeing an East Asian face may trigger a "forever foreigner" stereotype [11] and that the sociolinguistic expectations of East Asian and Caucasian faces can induce distinct perceptual responses.

An untested question is whether the perceived ethnicity of visual guises also impacts *speech production*. Prior work has shown that speakers tend to produce more effortful speech (e.g., through a slower speaking rate) when communicating with listeners who may have difficulty understanding them, such as non-native speakers or individuals who are hard-of-hearing [12]. These acoustic-phonetic modifications, also known as "clear speech", benefit listeners by increasing intelligibility compared to more "casual speech" [13]. Clear and casual speech are often elicited through explicit instructions to speakers (e.g., "Speak clearly to someone who may have trouble understanding you"; "Say the sentences in a natural, casual manner" [14]). However, it is not yet known whether the presentation of different visual guises (without written instructions) will affect speech production.

The current study tests three hypotheses. If seeing an Asian face triggers expectations that the listener is a non-native speaker with difficulty understanding English, then: (1) speakers should hyper-articulate (i.e., produce greater intensity, higher pitch, slower speaking rate, and greater pitch range) when looking at an Asian face compared to a Caucasian face; (2) speech patterns in the Asian and Caucasian face conditions should be similar to non-native-directed speech and casual speech, respectively (where the latter two styles are elicited via explicit instructions); (3) individuals who rate the Asian face as being more likely to be a non-native speaker relative to the Caucasian face should hyper-articulate more when looking at the Asian face.

2. METHODS

2.1. Participants

48 native English speakers (35 female, 13 male, 0 non-binary; mean age = 19.17; sd = 1.71) were recruited from the University of California, Davis (UC Davis) Psychology subjects pool and received course credit. One participant was removed for lack of attentiveness during the task.

2.2. Visual Guises

Figure 1 shows the images of the Asian and Caucasian faces. The pictures, which were selected from the Chicago Face Database, were rated as being females in their mid-20s and as conforming to their self-reported ethnicity in a norming study [15].

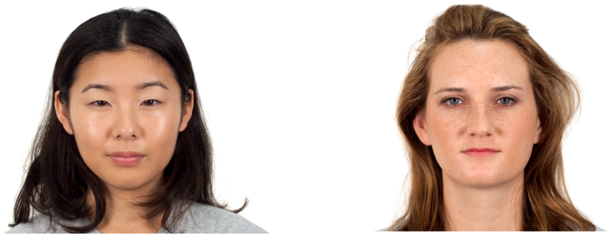

Figure 1: The Asian face (left) and the Caucasian face (right) used in this study.

2.3. Procedure

Participants completed two recording sessions on separate days between one and three days apart. In one session, there were two blocks – one presented the Caucasian face, and another presented the Asian face. The order of the blocks was counterbalanced. On each trial, speakers were shown a face and a sentence and asked to "produce the sentence to the listener in the image". After both blocks, participants rated the faces on how much they looked like non-native English speakers from 1 ("Strongly Disagree") to 7 ("Strongly Agree").

In another session, there were three blocks: (1) non-native-directed speech; (2) casual speech; (3) hard-of-hearing-directed speech (not analyzed here). To elicit non-native-directed speech, speakers were asked to imagine "talking to a listener who is a native speaker of Mandarin and is learning English". For casual speech, participants were asked to talk "casually" as if "to a listener who is a native speaker of English". The non-native-directed speech block always preceded the casual speech block. Neither block presented a visual image.

For all blocks, speakers produced the same 80 semantically unpredictable sentences (e.g., "Tom discussed the hay") from the Speech Perception in Noise Test [16]. Session order was counterbalanced. Productions were recorded in a sound-attenuated booth using a Shure WH20 XLR head-mounted microphone and digitally sampled at a 44-kHz rate.

2.4. Statistical Analysis

Intensity, pitch, speaking rate, and pitch range were measured over each sentence using Praat [17]. Speech rate was calculated by dividing the number of syllables by the sentence duration (in seconds). Each variable was modelled via separate linear mixed-effects regression models with the *lme4* R package [18]. All models had a treatment-coded fixed effect of Block (4 levels: Asian face, Caucasian face, Non-native-directed, Casual [reference level]) with by-listener and by-sentence random intercepts and by-listener random slopes for Block. The marginal and conditional coefficients of determination were found with the *r.squaredGLMM* function [19].

To test whether the ratings of the faces affected individual differences between the visual guise conditions, a correlational analysis was conducted. For each participant, the difference in ratings between the Asian and Caucasian faces was calculated, where a more positive difference indicates a greater perception that the Asian face corresponds to a non-native English speaker. For each participant, the difference between the mean value in the visual conditions was calculated for each acoustic variable. Pearson product-moment correlations were computed between the difference in ratings and the difference in the acoustic variables (e.g., testing if those who rated the Asian face as being more likely to be a non-native speaker also spoke with greater average intensity in the Asian face condition).

3. RESULTS

3.1. Intensity

Neither non-native-directed speech (*Coef* = 0.59, *SE* = 0.37, *z* = 1.62, *p* = 0.11), nor the Asian face condition (*Coef* = 0.11, *SE* = 0.57, *z* = 0.20, *p* = 0.84), nor the Caucasian face condition (*Coef* = 0.40, *SE* = 0.54, *z* = 0.74, *p* = 0.47) are significantly different in mean intensity from casual speech. The marginal and conditional coefficients of determination were 0.001 and 0.88, respectively.

3.2. Pitch

Neither non-native-directed speech (*Coef* = 1.86, *SE* = 1.10, *z* = 1.70, *p* = 0.10), nor the Asian face condition (*Coef* = -1.72, *SE* = 0.98, *z* = -1.77, *p* = 0.08), nor the Caucasian face condition (*Coef* = -1.44, *SE* = 2.97, *z* = -1.14, *p* = 0.26) are significantly different in mean pitch from casual speech. The marginal and conditional coefficients of determination were 0.004 and 0.81, respectively.

3.3. Pitch Range

Mean pitch range across conditions is provided in Figure 2. Non-native directed speech contains a significantly larger pitch range than casual speech (*Coef* = 8.12, *SE* = 1.32, *z* = 6.14, *p* < 0.001). There is no difference in pitch range between the Asian (*Coef* = 2.43, *SE* = 1.70, *z* = 1.43, *p* = 0.16) and Caucasian face conditions (*Coef* = 0.65, *SE* = 1.76, *z* = 0.37, *p* = 0.71) from casual speech. The marginal and conditional coefficients of determination were 0.008 and 0.65, respectively.

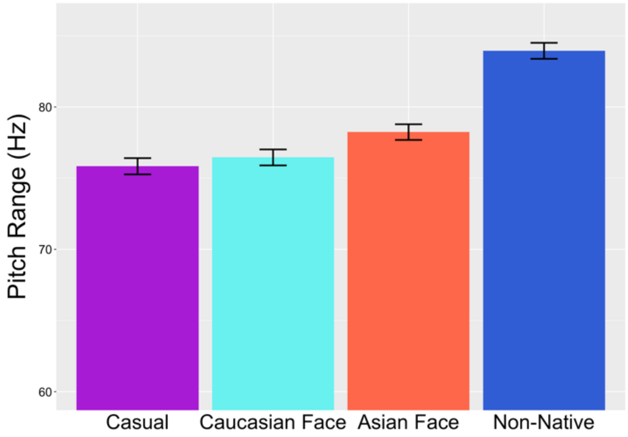

Figure 2: Mean pitch range (Hz) by condition. Error bars represent standard errors.

3.4. Speaking Rate

Figure 3 shows mean speaking rate by condition. Non-native-directed speech (*Coef* = -1.48, *SE* = 0.1, *z* = -14.75, *p* < 0.001), the Asian (*Coef* = -0.45, *SE* = 0.09, *z* = -4.81, *p* < 0.001), and Caucasian face conditions (*Coef* = -0.26, *SE* = 0.06, *z* = -4.14, *p* < 0.001) are all slower than casual speech. The marginal and conditional coefficients of determination were 0.30 and 0.85, respectively. A model with the same random effects structure that excluded casual speech and coded non-native-directed speech as the baseline level indicated that non-native directed speech is even slower than the Asian (*Coef* = 1.03, *SE* = 0.10, *z* = 10.54, *p* < 0.001) and Caucasian face conditions (*Coef* = 1.23, *SE* = 0.11, *z* = 11.35, *p* < 0.001). A final post-hoc model that only included the two visual guise conditions revealed that speaking rate in the Asian face condition is slower than in the Caucasian face condition (*Coef* = -0.20, *SE* = 0.08, *z* = -2.55, *p* = 0.01). The speaking rate results can be summarized as follows, from fastest to slowest: Casual > Caucasian Face > Asian Face > Non-Native-Directed.

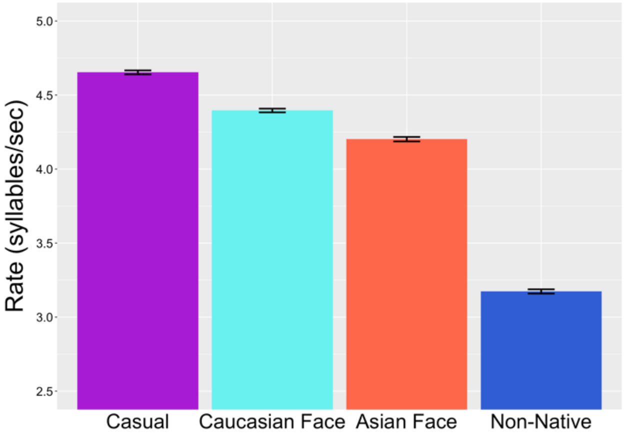

Figure 3: Mean speaking rate (syllables per second) by condition. Error bars represent standard errors.

3.5. Non-Native Speaker Ratings

The difference in ratings between the Asian and Caucasian faces did not correlate significantly with differences between the visual conditions for any acoustic variable (Intensity: *r* = -0.01, *p* = 0.97; Pitch: *r* = -0.22, *p* = 0.14; Pitch Range: *r* = -0.01, *p* = 0.95; Rate: *r* = -0.28, *p* = 0.053).

4. DISCUSSION

The current study supports the hypothesis that speakers hyper-articulate more when looking at a picture of an East Asian face compared to a picture of a Caucasian face. More specifically, speakers talk more slowly towards an East Asian face than a Caucasian face. However, speech towards an imagined non-native listener, which is hyper-articulated compared to casual speech through greater pitch range and slower speaking rate, is even slower than the Asian face condition. We also find that explicit ratings of how much the faces correspond to non-native English speakers do not correlate with differences in production between the visual conditions.

Overall, the slower speaking rate in the East Asian face condition relative to the Caucasian face condition is in line with a considerable body of work in speech perception on effects of visual guise [5, 6, 7, 8, 9, 10]. Visual cues, such as apparent ethnicity, can trigger expectations about group membership based on linguistic stereotypes. The assumption that East Asians are non-native English speakers leads to downstream effects in both speech perception and production. Similar to how listeners demonstrate better understanding of Mandarin-accented speech with a congruent East Asian face [9], speakers talk more slowly towards an East Asian face to aid listener comprehension. The current results also align with prior work in syntax and pragmatics indicating that listeners accommodate non-native speakers to facilitate communication (e.g., through greater leniency for syntactic errors [20] and for pragmatically odd or under-informative utterances [21, 22]).

Although speakers talk more effortfully towards an East Asian face than a Caucasian face, even greater hyper-articulation is observed in non-native-directed speech on pitch range and speaking rate. One way to account for the larger effect size in the non-native-directed speech condition is by appealing to Hypo- and Hyper-articulation (H&H) Theory [23] and the concept of *informativity* from Bayesian frameworks of speech perception [24, 25]. According to H&H Theory, articulatory patterns reflect competing goals between conserving effort and maximizing clarity. Whether speakers choose to produce more hyper-articulated speech, at the expense of articulatory effort, depends on how likely they think the listener will understand them. This likelihood is partially determined based on the informativity of socio-indexical cues. For example, in the non-native-directed speech condition in the current study, speakers are given an explicit description of the imagined listener as a "native speaker of Mandarin" who is "learning English". This written guise provides high certainty about the language background of the listener and is thus an informative cue. Given prior experience interacting with non-native listeners, speakers may talk with a highly effortful style to maximize clarity and reduce potential comprehension difficulties. In contrast to a written guise, visual cues to apparent ethnicity are less informative. Although ethnicity and language background can be correlated [26], they are not in a one-to-one relationship [27]. In fact, the majority of Asian Americans in the United States are actually proficient in English [28]. Visual guises thus offer less information about the language background of the listener, leading to greater uncertainty about whether the listener is a native or a non-native speaker, and therefore, whether hyperarticulation is necessary to facilitate communication. While linguistic stereotypes and biases may lead to a slower speaking rate towards East Asian faces than Caucasian faces, uncertainty about listener identity means that speakers are less willing to expend articulatory effort, thereby resulting in a smaller effect size compared to the non-native-directed condition.

Although speech towards an East Asian face is hyper-articulated compared to a Caucasian face, both visual conditions are significantly slower compared to casual speech towards an imagined listener who is a native English speaker. One possible explanation is that, relative to speech directed towards an imagined listener, visual guises simulate a more ecologically valid interaction with a real listener, thus resulting in a distinct speaking style. This is consistent with prior work showing acoustic differences between speech towards a real listener and speech towards an imagined listener [29, 30, 31]. Future work should more explicitly test how real-listener-directed speech compares acoustically to guise-directed speech.

Another direction for future research is to examine the role of implicit biases on speech towards visual guises. In the current study, speakers were asked to provide explicit ratings for the East Asian and Caucasian faces on how likely they corresponded to a native or non-native speaker, and no correlations were found between the ratings and the acoustic results. Prior studies, however, have shown that explicit judgments and implicit beliefs do not necessarily overlap [32]. More work is needed to see whether more implicit measures, such as the Implicit Association Test, can better account for individual differences in production towards visual guises.

Speaking style could also differ between other pairs of guises, such as a picture of a digital device and a human. Recent studies have proposed that, similar to East Asian individuals, people may be biased towards thinking that voice-activated artificially intelligent assistants have difficulty understanding them [33, 34, 35]. Consistent with these accounts, Aoki, Cohn, and Zellou [36] showed that presenting a picture of a cylindrical device decreases transcription accuracy in a speech-perception-in-noise task compared to a picture of a human for both a naturally produced voice and a text-to-speech voice. Bias against digital devices may elicit a more hyper-articulated speaking style towards a picture of a device compared to a human.

5. CONCLUSION

Although much past work has examined how the perceived ethnicity of visual guises affects speech perception, little work has explored effects on speech production. The current study begins to fill this gap, showing that speakers hyper-articulate towards East Asian faces compared to Caucasian faces through a slower speaking rate. These results lead to further questions about the nexus between speech perception, speech production, and social expectations.

6. REFERENCES

[1] Zheng, Y., Samuel, A. G. 2017. Does seeing an Asian face make speech sound more accented? *Atten. Percept. Psychophys.* 79, 1841-1859.

[2] Hanulíková, A. 2018. The effect of perceived ethnicity on spoken text comprehension under clear and adverse listening conditions. *Linguist. Vanguard* 4, 1-9.

[3] Hanulíková, A. 2021. Do faces speak volumes? Social expectations in speech comprehension and evaluation across three age groups. *PloS One*, 16, e0259230.

[4] Melguy, Y. V., Johnson, K. 2021. General adaptation to accented English: Speech intelligibility unaffected by perceived source of non-native accent. *J. Acoust. Soc. Am.* 149, 2602-2614.

[5] Rubin, D. L. 1992. Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Res. High. Educ.* 33, 511-531.

[6] Kang, O., Rubin, D. L. 2009. Reverse Linguistic Stereotyping: Measuring the Effect of Listener Expectations on Speech Evaluation. *J. Lang. Soc. Psychol.* 28, 441-456.

[7] Yi, H.-G., Phelps, J. E. B., Smiljanić, R., Chandrasekaran, B. 2013. Reduced efficiency of audiovisual integration for nonnative speech. *J. Acoust. Soc. Am.* 134, EL387-EL393.

[8] Hu, G., Su, J. 2015. The effect of native/non-native information on non-native listeners' comprehension. *Lang. Aware.* 24, 273-281.

[9] McGowan, K. B. 2015. Social Expectation Improves Speech Perception in Noise. *Lang. Speech.* 58, 502-521.

[10] Babel, M., Russell, J. 2015. Expectations and speech intelligibility. *J. Acoust. Soc. Am.* 137, 2823-2833.

[11] Tuan, M. 1998. *Forever foreigners or honorary whites?: the Asian ethnic experience today.* Rutgers University Press.

[12] Uchanski, R. M. 2005. Clear speech. In: Pisoni, D. B., Remez, R. E. (eds), *The Handbook of Speech Perception.* Blackwell, 207-235.

[13] Smiljanić, R., Bradlow, A. R. 2009. Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Lang. Linguist. Compass.* 3, 236-264.

[14] Cohn, M., Pycha, A., Zellou, G. 2021. Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech. *Cognition* 210, 104570.

[15] Ma, D. S., Correll, J., Wittenbrink, B. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* 47, 1122-1135.

[16] Kalikow, D. N., Stevens, K. N., Elliott, L. L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.* 61, 1337-1351.

[17] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 341-345.

[18] Bates, D., Mächler, M., Bolker, B. M., Walker, S. C. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1-48.

[19] Nakagawa, S., Johnson, P. C. D., Schielzeth, H. 2017. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface* 14, 20170213.

[20] Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., Weber, A. 2012. When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *J. Cogn. Neurosci.* 24, 878-887.

[21] Gibson, E., Tan, C., Futrell, R., Mahowald, R., Konieczny, L., Hemforth, B., Fedorenko, E. 2017. Don't underestimate the benefits of being misunderstood. *Psychol. Sci.* 28, 703-712.

[22] Fairchild, S., Papafragou, A. 2018. Sins of omission are more likely to be forgiven in non-native speakers. *Cognition* 181, 80-92.

[23] Lindblom, B. 1990. Explaning phonetic variation: A sketch of the H&H theory. In: Hardcastle, W. J., Marchal, A. (eds), *Speech production and speech modelling.* Springer, 403-439.

[24] Kleinschmidt, D. F., Jaeger, T. F. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev.* 122, 148-203.

[25] Kleinschmidt, D. F. 2019. Structure in talker variability: How much is there and how much can it help? *Lang. Cogn. Neurosci.* 34, 43-68

[26] Wolfram., W., Schilling, N. 2015. *American English: Dialects and Variation.* John Wiley & Sons.

[27] King, S. 2020. From African American vernacular English to African American language: Rethinking the study of race and language in African American's Speech. *Annu. Rev. Linguist.* 6, 285-300.

[28] Budiman, A., Ruiz, N. G. 2021. Key facts about Asian Americans, a diverse and growing population. Pew Research Center. https://www.pewresearch.org/short-reads/2021/04/29/key-facts-about-asian-americans/ (accessed Apr. 24, 2022).

[29] Scarborough, R., Bernier, J., Zhao, Y., Hall-Lew, L., Dmitrieva, O. 2007. An Acoustic Study of Real and Imagined Foreigner-Directed Speech. *Proc. 16th ICPhS* Saarbrücken, 2165-2168.

[30] Scarborough, R., Zellou, G. 2013. Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *J. Acoust. Soc. Am.* 134, 3793-3807.

[31] Scarborough, R., Zellou, G. 2022. Out of sight, out of mind: The influence of communicative load and phonological neighborhood density on phonetic variation in real listener-directed speech. *J. Acoust. Soc. Am.* 151, 577-586.

[32] Devos, T., Banaji, M. R. 2005. American = white? *J. Pers. Soc. Psychol.* 88, 447-466.

[33] Cohn, M., Liang, K. H., Sarian, M., Zellou, G., Zhou, Y. 2021. Speech Rate Adjustments in Conversations With an Amazon Alexa Socialbot. *Front. Commun.* 6, 671429.

[34] Cohn, M., Zellou, G. 2021. Prosodic Differences in Human- and Alexa-Directed Speech, but Similar Local Intelligibility Adjustments. *Front. Commun.* 6, 675704.

[35] Cohn, M., Segedin, B. F., Zellou, G. 2022. Acoustic-phonetic properties of Siri- and human-directed speech. *J. Phon.* 90, 101123.

[36] Aoki, N. B., Cohn, M., Zellou, G. 2022. The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise. *JASA Express Letters* 2, 045204.