Note. This version is subject to peer review of Speech Prosody as of January 7th, 2024.
It was uploaded to https://wenjunchen29.github.io/web/ on January 12th, 2024.

Inconsistent prosodies more severely impair speaker discrimination of Artificial-Intelligence-cloned than human talkers

Wenjun Chen[1], *Xiaoming Jiang*[12], *Jingyi Ge*[1], *Shuwan Shan*[1], *Siyuan Zou*[1], *Yiyang Ding*[3]

[1]Institute of Linguistics, Shanghai International Studies University, Shanghai 201620, China
[2]Key Laboratory of Language Science and Multilingual Intelligence Applications, Shanghai International Studies University, Shanghai 201620, China
[3]School of Russian and Eurasian Studies, Shanghai International Studies University, Shanghai 201620, China

wenjun.chen@shisu.edu.cn, xiaoming.jiang@shisu.edu.cn

Abstract

AI algorithms designed to clone human speaker identity are reportedly capable of replicating human-specific vocal expression. However, whether listeners can identify a single speaker expressing varying emotive states as the same individual remains unclear, particularly never in AI-to-AI pairings. This study asked thirty-six Chinese participants to judge whether identical speakers delivered pairs of Chinese sentences with incongruent or congruent prosody in human-only and AI-only scenarios. We found a decrease in the accuracy of identifying the same speaker under inconsistent prosody conditions compared to consistent ones, a trend evident in both human-to-human and AI-to-AI pairs. Meanwhile, correctly distinguishing between two speakers was more challenging than identifying a single speaker, with AI pairs reporting notably poorer performance than human-human pairs. When presented with pairs of speakers using consistent prosody, listeners demonstrated significantly slower reaction times when identifying two speakers. Our findings suggest that vocal prosodies can lead to within-speaker identity variation, in which listeners form average-based representations and still recognise the same speaker across prosodies. The findings about the reduced capability in speaker discrimination in AI voices provide supportive evidence for the "out-group homogeneity effect" of AI voice perception.

Index Terms: voice cloning, speech synthesis, vocal confidence, speaker discrimination, voice identity

1. Introduction

Human voice conveys both short-term emotional states and long-term information like age and speaker identity, and such paralinguistic information can be decoded through computational models and human listeners [1, 2]. Notably, such paralinguistic cues can also be encoded by synthetic talkers, and listeners differently perceive the human and synthetic talkers in dimensions like truthfulness and powerfulness in terms of person perception as well as softness, squeakiness, slowness, nasality and liveliness in regard to speech qualities [3]. Recent studies have suggested a human emotional intimacy effect where audiobook users find human-narrated speech more enjoyable and can better attract their attention and arouse more positive emotional responses [4] – which could be attributed to synthetic algorithms' inability to express human-specific speech prosodies. Our previous work has noted that Huawei's Xiaoyi, initially used for cloning speaker identity, is capable of capturing the vocal confidence-related prosodic features (confident, doubtful, and neutral-intending). Applying such to its AI-cloned speaker models allows producing post-clone AI talkers that share both speaker identity and vocal confidence with the original human talkers [5].

Encoding speaker identity and expression in the voice could share some vocal physiological foundations. This can be evidenced by fundamental frequency (F0) and vocal tract length (VTL), which have been linked to how speech conveys the speakers' relatively stable characteristics, such as age, biological sex, and the specific identity – *who is talking* [6]. Meanwhile, the representation of speaker identity is largely subject to the short-term emotive states that lead to systematic VTL and F0 modulation; for example, the speaker extends VTL and lowers F0 to express confidently [5, 7].

Lavan et al. (2019) shifted speakers' VTL and F0 in the X-axis and Y-axis with *Praat* exposed listeners to learning speaker identities away from the centre and tested if listeners recognised the never-heard voices in the centre as 'old'. The experiment found listeners classified the never-heard voices as familiar, indicating the existence of an average-based representation of speaker identity [6]. Still, this study's VTL and F0 manipulation was not associated with prosodic- or context-specific pragmatic intentions, such as fear. Xu and Armony (2021) designed a similar training-testing task that exposed listeners to three talkers' identities in either neutral or fearful prosody in the training stage and tested if listeners could recognise the trained known speakers in one prosody still as 'old' when the speakers are expressing themselves in another prosody in the testing stage. They reported only an accuracy lower than but close to the chance level [8]. This study seemed to suggest that if VTL and F0 modulations are associated with pragmatic intentions, i.e., fearful vs. neutral prosodies, listeners could not recognise the same talker's identity in different prosodies.

Against this background, we hypothesise that speech-prosody-led VTL and F0 modulation (not highly expressive ones [9]) could shift the speaker's identity but within a range so that listeners could still recognise talkers in different prosodies as the same talker. To test this, we employed the AX discrimination rather than the previous training-testing task, as the AX discrimination task can be more suitable for identity comparison and recognition in adverse conditions, such as reversed speech in the mother tongue and not familiar language [10]. Our study employed a design with 2 prosodies (confident vs. doubtful) * 2 sources (pre-clone human vs. post-clone AI) * 2 prosody pairs (inconsistent vs. consistent) * speaker pairs 2 (same vs. different) design. Listeners were tasked to decide whether a pair of voices was produced by the same speaker or not, ignoring the other differences between the sounds, such as prosodies.

2. Speaker discrimination study

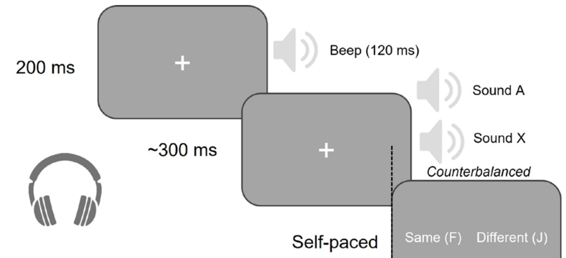
2.1. Methods

2.1.1. Participants

Thirty-six native Mandarin Chinese speakers, university students (26 females/10 males; Mean ± SD Age: 20.52 ± 2.62 years for females, 21.00 ± 2.72 years for males; Years of Education: 17.04 ± 2.13 for females, 17.55 ± 1.97 for males) without reported auditory or mental impairments participated in the perception experiment. Compensation was set at 50 RMB per hour. The experiment was approved by the Ethics Committee of the Institute of Linguistics, Shanghai International Studies University.

2.1.2. Material and paradigm

The auditory stimuli were selected from a larger audio corpus (24 speakers). Part of the preliminary validation of this audio corpus (10 speakers) was reported in our recent work [5]. The audio selected for this experiment ensured that for the same sentence, for instance, *I can fill in the form for you* (in Chinese), the AI-cloned voice and the human-produced voice for both confident and doubtful prosodies were perceptually distinct in confidence level, with the confident condition scoring higher than the doubtful one. A total of 24 speakers were paired based on biological sex to form 12 contrasting pairs, also ensuring close similarity in speaker height within each pair.



Source Type	Prosody Type	Prosody Consistency	Speaker
Human	Confident	IP	One
Human	Confident	CP	One
Human	Doubtful	IP	One
Human	Doubtful	CP	One
Cloned-AI	Confident	IP	Two
Cloned-AI	Confident	CP	Two
Cloned-AI	Doubtful	IP	Two
Cloned-AI	Doubtful	CP	Two

Note. Prosody Consistency and Speaker are designed according to Sound X, which is not illustrated in this table but can be deduced.

Figure 1: Study design of the AX Discrimination Task. IP is for Inconsistent Prosody, and CP is for Consistent Prosody.

Listeners were presented with a beep sound followed by two consecutive audio clips, namely Sound A and Sound X, and were required to judge whether the same person spoke the two sentences by pressing keys (F or J). The assignment of F and J to responses was counterbalanced, as was the order of AX sentences across participants. Participants were exposed to the design of *Prosody Consistency*: Consistent Prosody (CP) and Inconsistent Prosody (IP). The pairs of sounds were from the same source with both human voice and AI-cloned voice. A Latin square design was employed to counterbalance the block order among participants. See Figure 1.

2.1.3. Data analysis

The Effects of Talker and Prosody Consistency on Accuracy. With the *lme4* package (Version 1.1-35.1)[11] on RStudio version 2023.09.1 Build 494, our initial Mixed-effects logistic regression model (Model 1-a) utilised the formula of *Response Accuracy ~ Prosody Consistency (CP and IP) * Talker (One and Two) + (1 | Participant), family = binomial*. For '*Talker (One and Two)*', the response accuracy corresponds to the correct answer for the pair: same talker – one talker and different talkers – two talkers. The main effect and interaction results are documented in Table 1. Significant post hoc results are annotated in Figure 2 (contrasting one vs. two speakers in CP-CP or IP-IP).

We further included speaker identity sources into the model (Model 1-B) with a formula: Response Accuracy ~ Prosody Consistency (CP and IP) * Talker (One and Two) * Sources (human vs. AI) + (1 | Participant), family = binomial in a second fitting. The reason for adding Sources is that we suspect the contexts of AI-AI and human-human pairs could have an interaction effect with other factors. The main effect and interaction results have been reported above. Significant post hoc results are reported in Table 2, annotated in Figure 2 (contrasting AI and human trials in each sub-comparison).

The Effects of Talker and Prosody Consistency on Reaction Time (RT). With lme4, we first fitted a smaller LMER model (Model 2-A): Reaction Time ~ Prosody Consistency (CP and IP) * Talker (One and Two) + (1 | Participant). Significant post hoc results are annotated in Figure 3 (contrasting one vs. two speakers in CP-CP or IP-IP).

We then fitted another larger model (Model 2-B) with a formula of Reaction time ~ Prosody Consistency (CP and IP) * Talker (One and Two) * Sources (human vs. AI) + (1 | Participant). Sources are added to explore if they interact with other factors. Significant post hoc results are reported in Table 4, annotated in Figure 3 (contrasting AI and human trials in each sub-comparison).

3. Results

3.1. Prosody Consistency and Talker Sources on Accuracy

For Model 1-A, prosody consistency and the number of speakers influenced participants' likelihood of correctly identifying paired stimuli within human- or AI-only contexts, as shown in Table 1 for Model 1-A.

The IP level of the *Prosody Consistency* variable, relative to the CP baseline, revealed a substantial effect. An β of -2.64 implied lower log odds for the IP condition compared to CP. The Odds Ratio of .07 indicated that the chance of a correct answer under the IP was 7% of that under CP, suggesting IP had a lower accuracy than CP.

In scenarios requiring participants to differentiate between voices from two distinct talkers, as opposed to a single speaker, the model suggested a notable decrease in the likelihood of a correct response. An β of -3.08 indicated a significant reduction in accuracy with two speakers. The corresponding Odds Ratio of .05 indicated that the probability of a correct response with two speakers was merely 5% of the likelihood with one speaker, highlighting the tendency of perceiving two speakers as one.

Note. This version is subject to peer review of Speech Prosody as of January 7th, 2024.
It was uploaded to https://wenjunchen29.github.io/web/ on January 12th, 2024.

The interaction effect (β = 3.08) in the IP condition with two speakers significantly increased the odds of a correct response compared to one speaker in the CP baseline. However, this does not mean the overall accuracy for IP-Two was necessarily higher than CP-One. This is because of the negative independent effect of 'Two' (β = -3.08) which suggested that discriminating between more speakers generally lower response accuracy. Thus, while the interaction suggested a relative improvement under IP with two speakers, the actual overall accuracy rate depended on the combined effects of all factors, not solely on the interaction.

Table 1: Mixed-Effects Logistic Regression Results for Model 1-a

Term[a]	β	**SE**	**z**	*p*[b]	**OR**[c]	**95% CI**
Intercept	4.33	.22	20	***	76.11	[50.85-119.65]
IP	-2.64	.22	-11.99	***	.07	[.04-.11]
Two	-3.08	.21	-14.42	***	.05	[.03-.07]
IP:Two	3.08	.23	13.22	***	21.75	[14.01-35.16]

[a] IP (from IP & CP); Two (from One & Two Talkers)

[b] Significance codes: $p < .001$***, $p < .01$**, $p < 0.05$*

[c] The 'Odds Ratio' column is the exponentiated version of the estimates, which gives the change in odds for a one-unit increase in the predictor variable.

For Model 1-B (Table 2 and Figure 2), the newly added *Sources (human vs. AI* were found to have no main effect (p=.22), no interaction with *Prosody Consistency* (p=.15) and *Talker* (p=.55), and no three-level interaction with the other two (p=.14).

Still, we observed AI-AI and human-human pairs could influence listeners' performance in detailed one vs. two talker-discrimination tasks through post hoc analysis. The results suggested that listeners' performance of accurately discriminating between two talkers was higher in human-human pairs than in AI-AI pairs.

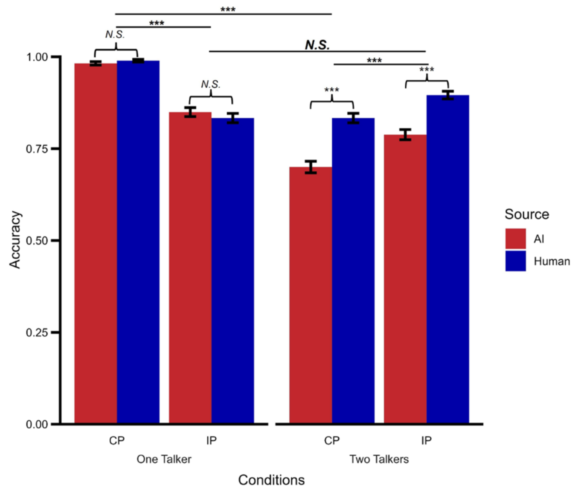


Figure 2: Talker numbers, prosody consistency, and speakers' group identity influence accuracy.

Table 2: Post Hoc Results for Model 1-B

Contrast[a]	β	**SE**	**z**	*p*[b]
One CP AI - One IP AI	2.36	.28	8.39	***
One CP AI - One CP Human	-.52	.42	-1.23	.92
One IP AI - One IP Human	.12	.13	.93	.98
Two CP AI - Two IP AI	-.42	.12	-3.43	*
Two CP AI - Two CP Human	-.78	.12	-6.48	***
Two IP AI - Two IP Human	-.85	.14	-6.09	***
One CP Human - One IP Human	3.00	.35	8.58	***
Two CP Human - Two IP Human	-.49	.15	-3.21	*

[a] One CP AI: one AI speaker talking in consistent prosody in the pair.

[b] Significance codes: $p < .001$***, $p < .01$**, $p < .05$*

3.2. Prosody Consistency and Talker Sources on RT

Analysis of Model 2-A revealed that the sources, whether AI or human, had no statistically significant main effect on participants' Reaction times (F(1, 2280.4) = .93, p=.34).

The talker numbers had a main effect, suggesting that participants' reaction times varied depending on whether there was one or two talkers present (F(1, 6776.1)=43.64, $p < .001$, ηp^2= 6.40e-03). Post hoc contrast suggested that listeners reacted significantly slower when there were two talkers talking (One-Two: β=-.23, SE=.03, z=-6.61, p<.0001).

There was a significant interaction between the sound source and the number of talkers (F(1,6776.1)=22.67, $p < .001$, ηp^2= 3.33e-03), indicating that the effect of the number of talkers on Reaction time was influenced by whether the sound source was AI or human. This indicated that listeners' reaction times under conditions of one or two speakers were differentially affected by whether it was an AI-AI pair or a human-human pair.

Additionally, following a post hoc analysis, we found listeners reacted significantly slower in the two-talkers condition only in IP rather than CP (CP One - CP One: β=-.21, SE=.06, z=-3.74, p=.001).

For Model 2-B, the post hoc analyses, as indicated by the annotated signs in Figure 3 and results in Table 3, demonstrated that no condition displayed a significant difference. We observed neither the main effect of the newly added human vs. AI sources (p=.54) nor its interaction with *Prosody Consistency* (p=.78) and *Talker* (p=.19), nor three-level interaction (p=.99). Further, post hoc analysis contrasting AI-AI and human-human pairs did not report any significance for CP-One (p=1.0), IP-One (p=1.0), CP-Two (p=.96), and IP-Two (p=1.0).

Despite this, listeners were seemingly faster in AI pairs than human pairs only in one-speaker conditions, whereas they were slower in two-speaker conditions. However, this was supported only by visualisation (p>.005).

Table 3: *Post Hoc Results for Model 2-B*

Contrast[a]	β	**SE**	**z**	***p***
One CP AI - One IP AI	-.2	.08	-2.68	.13
One CP AI - One CP Human	-.02	.07	-.22	1
One IP AI - One IP Human	-.03	.07	-.48	1
Two CP AI - Two IP AI	.14	.08	1.81	.61
Two CP AI - Two CP Human	.08	.07	1.1	.96

Two IP AI - Two IP Human	.06	.07	.83	.99
One CP Human - One IP Human	-.22	.08	-2.92	.07
Two CP Human - Two IP Human	.12	.08	1.54	.79

[a] One CP AI: one AI speaker talking in consistent prosody in the pair.

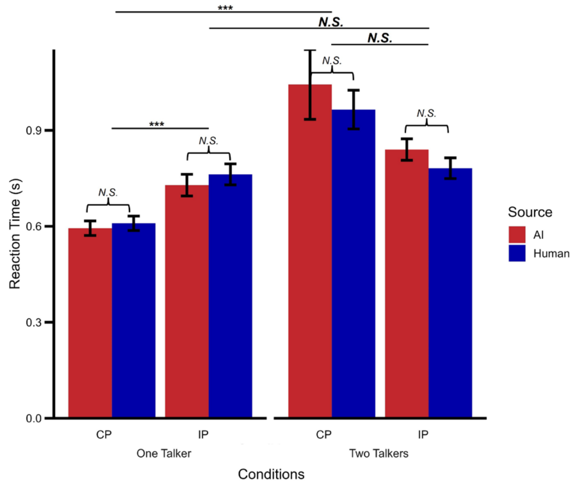


Figure 3: Talker numbers, prosody consistency, and speakers' group identity influence reaction time.

4. Discussion

Our results demonstrate that listeners' accuracy and speech in correctly identifying whether two audio segments belong to the same speaker or rejecting them as two different people are related to the consistency of prosody in successive segments and whether the audios are AI-generated or human pairs.

Firstly, we observed that the accuracy of listeners in identifying the same speaker as one talker significantly decreases with inconsistent prosody compared to consistent prosody. This pattern, highlighting the influence of prosody inconsistency on speaker discrimination, is similar to previous results in old-new judgments of fearful and neutral audio, but it also extended previous studies by Xu and Armony (2021), which (more of its experiment design see the Introduction) suggested that listeners could only integrate and reject speakers with different prosodies at a chance level [8]. This improved integration performance might be due to the lower cognitive resources required by the AX discrimination task, which does not require listeners to carry forward speaker identity information for later recognition [10], unlike the training-testing task that demands listeners to become thoroughly familiar with individual speaker identities to adapt to the internal identity changes brought by audio VTL/F0 modulations [12] and increased vocal expressiveness [9]. Thus, the AX paradigm in the current study is suitable for directly exploring listeners' integration of speaker identity across speech prosody, similar to identifying talkers across languages [13]. The results also indicate that listeners can integrate speakers with different emotional states into the same identity far above the chance level, even though inconsistent prosody reduces accuracy and causes confusion. This result was consistent in both AI and human pairs, further hinting at the human-like nature of AI-generated audio and its perceived role in human listeners. Hence, our study suggests that while inconsistent prosody can complicate speaker recognition, listeners can still effectively integrate the same talker in different prosodies; the similar performance pattern in extracting speaker identity commonality in both AI and human speech suggests that AI-generated voices can be designed in ways that are familiar and easily interpretable by human listeners, potentially enhancing the effectiveness of AI in roles that require speech interaction with humans [14].

Secondly, it is noteworthy that listeners were more likely to correctly identify two speakers in IP than in CP. This might reflect a tendency: when prosody is inconsistent, the perceived distance between speaker identities is greater, e.g., confident prosody has an averagely longer VTL and shorter F0 than doubtful prosody [5, 7]. Lavan et al. (2019) concluded that when stimuli were acoustically closer to the centre or average of a voice identity's representation, the accuracy in identifying these voices increased [6]. Our studies suggest that two talkers' VTL and F0 are more distinct when they express different emotive states [5], leading listeners to reject them as two speakers more easily.

Thirdly, in cases where listeners needed to reject two speakers correctly as different individuals, performance in AI-AI conditions was significantly less accurate than in human-human conditions, in both IP and CP. This suggests that while being presented with two speakers, listeners are more inclined to consider AI speakers as the same individual, i.e., they tend to think that AI audios represent the same person, even if the setup involves two speakers. This tendency might be related to the categorisation perception or the out-group homogeneity effect, where people perceive members of an out-group as more similar to each other than members of their own group (in-group) [15]. Hence, less familiar or categorised groups (i.e., AI voices) are perceived as less distinct from each other.

Finally, our findings suggest that listeners need significantly more time to decide when presented with two speakers with consistent prosody, but this effect is not seen when prosody is inconsistent. In our current design, the reaction time was calculated from the end of the second audio segment, and this could lead to delayed detection of online speaker discrimination. After all, previous studies have suggested that speaker group identity can be differentiated as early as 100ms, as indicated by the N100 amplitude being sensitive to out-group attitudes [16]. We suppose the longer reaction time for telling two speakers apart can be related to the higher cognitive load required to process two different speaker identities. In the case of inconsistent prosody, the lack of significant reaction time differences between one and two talkers might be due to the inherently slower responses in this condition, possibly because processing prosody and distinguishing between speakers may occur in parallel rather than in an additive manner [16]. This hypothesis would benefit from support from time-sensitive data, taking the start of the second audio segment as the marker.

5. Acknowledgements

This work was supported by the Natural Science Foundation of China (Grant No. 31971037); the 'Shuguang Programme' supported by the Shanghai Education Development Foundation and Shanghai Municipal Education Committee (Grant No. 20SG31); the Natural Science Foundation of Shanghai (22ZR1460200); the Supervisor Guidance Programme of Shanghai International Studies University (2022113001); and the Major Programme of the National Social Science Foundation of China (Grant No. 18ZDA293).

6. References

[1] B. Schuller and A. Batliner, Computational paralinguistics: emotion, affect and personality in speech and language processing. John Wiley & Sons, 2013.

[2] M. Mileva and N. Lavan, "Trait impressions from voices are formed rapidly within 400 ms of exposure," *Journal of Experimental Psychology: General,* 2023.

[3] J. W. Mullennix, S. E. Stern, S. J. Wilson, and C.-l. Dyson, "Social perception of male and female computer synthesised speech," *Computers in Human Behavior* vol. 19, no. 4, pp. 407-424, 2003.

[4] E. Rodero and I. Lucas, "Synthetic versus human voices in audiobooks: The human emotional intimacy effect," *New Media & Society,* vol. 25, no. 7, pp. 1746-1764, 2023.

[5] W. Chen and X. Jiang, "Voice-Cloning Artificial-Intelligence Speakers Can Also Mimic Human-Specific Vocal Expression," in *Preprints*, ed: Preprints, 2023.

[6] N. Lavan, S. Knight, and C. McGettigan, "Listeners form average-based representations of individual voice identities," *Nature Communications,* vol. 10, no. 1, pp. 1-9, 2019.

[7] X. Jiang and M. D. Pell, "The sound of confidence and doubt," *Speech Communication,* vol. 88, pp. 106-126, 2017.

[8] H. Xu and J. L. Armony, "Influence of emotional prosody, content, and repetition on memory recognition of speaker identity," *Quarterly Journal of Experimental Psychology,* vol. 74, no. 7, pp. 1185-1201, 2021.

[9] N. Lavan, L. F. Burston, P. Ladwa, S. E. Merriman, S. Knight, and C. McGettigan, "Breaking voice identity perception: Expressive voices are more confusable for listeners," *Quarterly Journal of Experimental Psychology,* vol. 72, no. 9, pp. 2240-2248, 2019.

[10] D. Fleming, B. L. Giordano, R. Caldara, and P. Belin, "A language-familiarity effect for speaker discrimination without comprehension," *Proceedings of the National Academy of Sciences,* vol. 111, no. 38, pp. 13795-13798, 2014.

[11] D. Bates, "lme4: Linear mixed - effects models using Eigen and S4," *R package version,* vol. 1, p. 1, 2016.

[12] N. Lavan, S. Knight, and C. McGettigan, "Listeners form average-based representations of individual voice identities," *Nature Communications,* vol. 10, no. 1, p. 2404, 2019/06/03 2019.

[13] S. J. Winters, S. V. Levi, and D. B. Pisoni, "Identification and discrimination of bilingual talkers across languages," *The Journal of the Acoustical Society of America,* vol. 123, no. 6, pp. 4524-4538, 2008.

[14] G. Laban, A. Kappas, V. Morrison, and E. S. Cross, "Building Long-Term Human–Robot Relationships: Examining Disclosure, Perception and Well-Being Across Time," *International Journal of Social Robotics,* 2023.

[15] J. M. Ackerman *et al.*, "They all look the same to me (unless they're angry) from out-group homogeneity to out-group heterogeneity," *Psychological science,* vol. 17, no. 10, pp. 836-840, 2006.

[16] X. Jiang, K. Gossack-Keenan, and M. D. Pell, "To believe or not to believe? How voice and accent information in speech alter listener impressions of trust," *Quarterly Journal of Experimental Psychology,* vol. 73, no. 1, pp. 55-79, 2020.