VOICE DISCRIMINATION ACROSS SPEAKING STYLES IN PERSIAN

Elisa Pellegrino[1], Homa Asadi[2], Volker Dellwo[1]

[1]Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
[2]University of Isfahan, Isfahan, Iran

{elisa.pellegrino; volker.dellwo}@uzh.ch; h.asadi@fgn.ui.ac.ir

ABSTRACT

Human listeners have a remarkable ability to recognize speakers by their voice, but within-speaker voice variability through different speaking styles, for example, can reduce recognition performance. In this study, we investigated voice discrimination across speaking styles in Persian. One hundred and forty-three naïve Persian listeners were asked to decide whether pairs of style-matched utterances in child-directed, spontaneous, read and clear speech originated from the same or different speakers. Listeners' performance across speaking styles was examined using the bias-free sensitivity measure A', and the bias measure b''D. Results showed that listeners performed accurately across all speaking styles but more poorly in child-directed speech. They had a bias toward responding 'different' regardless of the speaking style, thus suggesting a more general difficulty telling people together than apart.

Keywords: speaker discrimination, speaking styles, vocal variability.

I. INTRODUCTION

Human voices are individual and convey numerous cues about talkers' identity, including age, gender, health conditions, or emotional states [1]. Beyond being an "auditory face" [2], human voices also act as a behavioural biometric [3]. However, they are susceptible to strong variability for the effect of spontaneous or intentional modulations [4]. Human voice changes, for example, across the life span [5], as an effect of fatigue [6] or physical and mental health [7], [8]. Still, we also adapt the way we speak according to the age of our interlocutors [[9], [10], their linguistic origin [11], as well as to the background conditions [12]. For a review of volitional and unvolitional changes see [13]. How human listeners can distinguish speakers despite the enormous variability that individual voices reveal is far from being fully understood. In this study, we investigated the effect of the variability introduced to voices by speaking styles on unfamiliar speaker discrimination. What are the factors affecting voice processing? It has been amply documented that human listeners are fairly accurate at distinguishing unfamiliar speakers, especially if exposed to long, phonetically and prosodically rich utterances [14]–[17]. Nevertheless, studies on speaker perception and forensic speaker identification have shown that linguistic and extra-linguistic sources of within-speaker variability challenge the perception of speaker identities. More specifically, it has been observed that listeners' accuracy to recognize or discriminate between unfamiliar voices is reduced by mismatch in the language spoken in the stimulus pairs [18][19], type of vocalizations (vowels vs laughter, [20]; sung vs spoken word [21]) and in the linguistic relationship between lexical items within-one stimulus pair [22]. Detrimental to speaker perception are also changes in degree of speakers' vocal expressiveness [23] and in voice quality obtained by disguising voice source or resonance features (e.g. creaky voice, falsetto, hyponasality, whispering [24] [25] [26].

Another factor affecting vocal identity processing, that has recently received more scholarly attention, is speaking style. Numerous studies have indeed focussed on the effect of style-matched and style mismatched conditions on the speaker discrimination abilities in humans and machines. Style matched trials typically featured read-read speech only [27] or both read-read and spontaneous-spontaneous speech [28], [29], [30]. Style mismatched trails, instead, combined read speech with spontaneous speech [27]–[30] or with pet-directed speech [31]. Overall findings from this line of research have revealed that both humans and machines performed better when the speaking style was coherent within one stimulus pair compared to when it changed. Moreover, between the style-matched conditions (read-read, spontaneous-spontaneous), the read–read condition scored the highest accuracy. Taken together, results suggest that less controlled speaking styles (e.g. conversational or free speech) and the exaggerated prosody of pet-directed speech introduce more variability in the speakers' acoustic spaces [29] that make it difficult for listeners to establish stable identity percepts. Lower accuracy in style mismatched as compared to matched automatic voice recognition, yet with novel

insights on the role of high variable speaking styles on speaker perception is also documented in [32]. In [31], speaking styles under comparisons were clear speech and conversational speech on the one hand, adult- and infant-directed speech on the other. Between the matched conditions, mean within-register recognition led to ceiling effects for adult-directed, clear and conversation speech. Slightly lower accuracy was documented for infant-directed speech. More interestingly, in the style mismatched conditions, it was observed that when the systems were trained with more acoustically variable speaking styles (spontaneous or infant-directed speech) and tested with less variable registers (clear speech or adult-directed speech), the system performed better as compared to the opposite conditions (training with clear or adult-directed speech and testing respectively with spontaneous and infant-directed speech) [32]. This suggests that — at least for machines — learning vocal identity from acoustically more variable speaking styles may also lead to speaker recognition benefits.

To shed further light on the effect of speaking styles on speaker perception, we tested the speaker discrimination performance of Persian listeners on stimuli produced in most of the speaking styles employed in previous research: read, clear, spontaneous and child-directed speech, this latter being comparable to pet directed speech in many relevant features [33]. In line with findings showing that listeners' performance degrades with the more inherently acoustically variable speaking styles ([28]–[31], we hypothesize that listeners are more accurate in clear and read speech than in spontaneous or infant-directed speech. If this hypothesis holds, we also expect listeners to be more likely to incorrectly assign intraspeaker variability to different individuals (telling people together) in spontaneous and child-directed speech, that should measurably result in a bias towards responding 'different'. Based on the novel observations that speaker-specific acoustic variability can instead benefit voice processing by machines [32] we cannot fully exclude that listeners may perform better in spontaneous and child-directed speech and exhibit a bias toward responding the same (e.g. incorrectly assign interspeaker variability to same individuals) in clear and adult directed speech.

2. MATERIALS AND METHOD

2.1. Speakers

52 male Persian speakers, aged between 25 and 45, y.o. (Mean=29.66 , SD=4.8) were recruited for the collection of the database of within-speaker speaking style variability. Participants were BA, MA or PhD students at the University of Isfahan. All participants were monolingual with no language other than Persian spoken in their homes. They were all from the Isfahan province where they had lived their entire lives. None of the speakers reported any history of speech or hearing disorders and they were all naïve as to the purposes of the experiment. Participants from the Department of Linguistics were rewarded a grade, while the others received monetary compensation.

2.2 Speech Materials

To study the effect of speaking style on speaker discriminability, we collected a speech corpus of child-directed, read, clear and spontaneous speech. For the read speech part, each speaker was instructed to read aloud a list of 20 sentences, at their natural pace and intonation, with a pause between each sentence. They were also told to repeat any disfluent sentences before moving on. For the clear speech part, the same set of sentences was used but participants were asked to imagine that they were talking to hard of hearing persons. Speakers had several attempts for this task, and we eventually selected the repetition of sentences which sounded more clearly enunciated. To elicit between 2-4 minutes of spontaneous speech, a set of topics was developed (e.g. study field, recent vacations, and plans for the future). Speakers were instructed to select one or two topics which they would be comfortable speaking about. To elicit child-directed speech, participants were shown a picture of a baby and asked to tell a story to him. The speaking styles together yielded between 7 to 9 minutes of speech per speaker for a total corpus of 416 minutes of speech.

Due to COVID restrictions, speech data collection was carried out remotely, in a quiet room at participants' home, via vocal messages through the WhatsApp application. To control for the effect of differences in mobile equipment in the quality of recording, the corpus was collected exclusively using one specific smartphone brand and model. Recording instructions, sentences and pictures were sent to the speakers prior to the recording session. To supervise the data collection procedure, the recordings were taken during video call on ZOOM in the presence of the experimenter (2nd author). Recording sessions for each speaker took place on the same day. To permit acoustic analysis in Praat, participant's audio tracks were converted from their original format supported by the device (.m4a) to .wav format using the free online app Online Audio Converter but the original sampling rate and bit depth (44 kHz; 32 bit) were not modified. Eighty audio clips of about 3 seconds were extracted from each speech samples of the dataset. Each excerpt was selected such that the influence of

semantic cues to speaking styles on listeners' responses was minimized.

2.2 Listeners

143 native Persian listeners (male=75, female=68), ranged in age from 18 to 36 y.o. participated in the speaker discrimination test. The listeners were all students at the University of Isfahan. None was a trained phonetician. All listeners declared not to have hearing, vision, and/or dyslexia problems.

2.3. Procedure

Listeners were tested individually in a quiet room at their home. Their task was to decide whether the two stimuli in a trial were spoken by the same speaker or by different speakers (voice discrimination task). Each stimulus in a stimulus pair was scaled to an average intensity of 70 dB SPL. The stimuli in each stimulus pair were separated by a 1s silent interstimulus interval. Each listener was presented with 40 stimulus pairs (10 per speaking styles) over high-quality earphones, equally divided by same and different speakers trials. The stimulus order and item pairs were randomised for each participant but the stimulus set was identical to all listeners. The pairing of voices in different-speaker trials was random. Experiment was designed through learning management system (LMS) of the University of Isfahan, which provided an accessible online platform for conducting behavioural studies. Listeners were shown the "same" and "different" options on the screen and asked to click on the corresponding button after listening to each trial. Listeners were familiarised with the experiment interface and the stimuli through a demonstration session which presented four random stimuli containing voices and lexical items not present in the stimulus set. Testing was self-paced; participants generally took approximately 5-8 minutes to complete the experiment. Participants used their in-built microphones and headphones when doing the experiment. The study was conducted within the guidelines of the Ethics Committee of the University of Isfahan. Participants gave their informed consent to participate in the study.

2.4. Data Analysis and Statistics

To analyse the effect of speaking styles on listeners' ability to discriminate between same and different speakers, we calculated the bias-free sensitivity measure A' from signal detection theory [34]. A' values range from 0.0 to 1.0, with 0.5 signifying chance level sensitivity and 1.0 indicating highest sensitivity. A' sensitivity measure (dependent variable was calculated per listener and speaking style. To examine the effect of speaking style on listeners bias towards responding the same or different, we also calculated the measure b''D per listeners and speaking styles. B''D scores range between 1 and -1, with negative values indicating a bias toward responding 'different speakers', and positive values 'same speaker'. To test the significance of the effect of speaking style on A' and b''D, we ran Linear Mixed Effect Model with Speaking Styles as a fixed factor, A' and b''D as dependent variables, listeners as random intercept. Statistical analyses were performed with R Core Team 2022 [35].

3. RESULTS

As shown in Fig. 1, speaker discrimination was performed accurately in all speaking styles, albeit to a different extent. The lowest sensitivity was scored in stimuli produced in child-directed speech followed, in increasing order of sensitivity, by read, spontaneous and clear speech. The results of statistical analysis confirmed in part these observations: the effect of speaking style on A'[χ^2 (3)=22.162, p <0.001] was significant. Post-hoc analysis with Tukey corrections was conducted that showed that only the comparisons including stimuli in child-directed speech were significant (Fig. 2). The negative sign of the estimate indicated a poorer performance in child-directed speech as compared to clear, read and spontaneous speech. Concerning the bias, Fig. 3 shows that the scores of b''D were mostly negative for all the examined speaking styles, thus pointing to a more general bias towards responding 'different'. This pattern was confirmed by the statistical analysis that revealed no significant changes in b''D between read, clear, spontaneous and child-directed speech [χ^2(3)=6.7, p = 0.082].

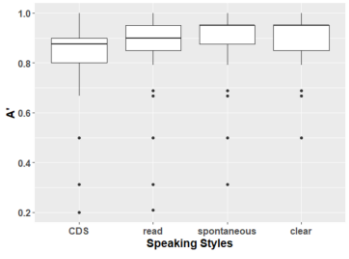

Figure 1: A' values by Speaking Style. The styles in the plot are ordered by increasing listeners' sensitivity

```
$contrasts
 contrast           estimate     SE  df t.ratio p.value
 CDS - clear       -0.05490 0.0127 435  -4.333  0.0001
 CDS - read        -0.03683 0.0127 435  -2.906  0.0200
 CDS - spontaneous -0.04872 0.0127 435  -3.845  0.0008
 clear - read       0.01808 0.0127 435   1.427  0.4833
 clear - spontaneous 0.00618 0.0127 435   0.488  0.9618
 read - spontaneous -0.01190 0.0127 435  -0.939  0.7839

Degrees-of-freedom method: kenward-roger
P value adjustment: tukey method for comparing a family of 4 estimates
```

Figure 2: Results of post-hoc test with Tukey correction for multiple comparisons across speaking styles.

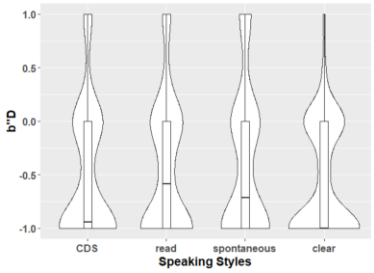

Figure 3: b''D values by Speaking Style.

5. DISCUSSION

In this study, we investigated the effect of speaking styles on Persian listeners' ability to discriminate between speakers. We quantified listeners' performance in terms of the bias-free sensitivity measure A' and the bias measure b''D. The overall results of this study confirm that speaking style variability affects the speaker discrimination performance. The direction of such effect, i.e. the lower sensitivity in child-directed speech, supports previous research indicating that inherently more variable speaking styles pose challenges to speaker discrimination [28]–[31]. Surprisingly though, spontaneous speech did not worsen listeners' performance compared to read speech as expected and reported in previous research [28]–[30]. No significant differences in A' were, indeed, obtained between clear, spontaneous and read speech. What causes this discrepancy is unclear and will be object of future acoustic investigations. Here we can only speculate that with the controlled elicitation method used for this study we may have not obtained a clear acoustic divide between clear, spontaneous and read speech, which results in measurably comparable difficulty when trying to discriminate between unfamiliar speakers. Alternatively, individual speakers may have produced the four speaking styles differently, with some of them marking the inter-style differences to a higher degree as compared to others (cf. [36] for individual differences in spontaneous and read speech). Within-speaker acoustic variability across speaking styles draws paths for future acoustic investigations.

Another unpredicted finding of this study was the listeners' bias towards responding different, irrespective of speaking styles. This type of bias was expected only for the more inherently variable speaking styles of the corpus but not for clear and read speech. Such bias indicates that participants were more likely to incorrectly assign intraspeaker variability to different individuals than they were to incorrectly assign inter-speaker variability to the same individual. We propose here two alternative non-exclusive explanations for this finding. First, one may impute the observed bias to the cautious behaviour of our listeners that - in line with previous research with not degraded stimuli nor embedded in background noise [26] seem to prefer responding different when unsure about the origin of intra-stimulus voices (for different results with degraded stimuli see [26]). This result also echoes recent research showing that listeners have more difficulties telling people together than apart (see a.o. [12]). The alternative interpretation is that the stimuli used in the perception contain comparable within-speaker variability that induced listeners to mistake within for between-speaker variability. Because the acoustic signal is the input to human perceptual processes, future step of this research will examine the acoustic properties in the stimuli used in the perception test to understand the perceptual strategies used by Persian listeners when discriminating between unfamiliar speakers in different speaking styles.

6. ACKNOWLEDGEMENTS

This study was supported by a Swiss Research Partnership Grant with South Asia and Iran funded by the Zürcher Hochschule für Angewandte Wissenschaften (ZHAW).

7. REFERENCES

[1] V. Dellwo, M. Huckvale, and M. Ashby, "How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification," in *Speaker Classification I*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–20.

[2] P. Belin, S. Fecteau, and C. Bédard, "Thinking the voice: neural correlates of voice perception," *Trends Cogn Sci*, vol. 8, no. 3, pp. 129–135, Mar. 2004.

[3] S. J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P. A. Keating, and A. Alwan, "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," *J Acoust Soc Am*, vol. 144, no. 1, pp. 375–386, Jul. 2018, doi: 10.1121/1.5045323.

[4] Y. Lee and J. Kreiman, "Acoustic voice variation in spontaneous speech," *J Acoust Soc Am*, vol. 151, no. 5, pp. 3462–3472, May 2022, doi: 10.1121/10.0011471.

[5] B. V. Tucker, C. Ford, and S. Hedges, "Speech aging: Production and perception," *WIREs Cognitive Science*, vol. 12, no. 5, Sep. 2021, doi: 10.1002/wcs.1557.

[6] N. R. Williams, "Occupational groups at risk of voice disorders: a review of the literature," *Occup Med (Chic Ill)*, vol. 53, no. 7, pp. 456–460, Oct. 2003, doi: 10.1093/occmed/kqg113.

[7] H. Steurer, E. Schalling, E. Franzén, and F. Albrecht, "Characterization of Mild and Moderate Dysarthria in Parkinson's Disease: Behavioral Measures and Neural Correlates," *Front Aging Neurosci*, vol. 14, May 2022.

[8] J. Wang, L. Zhang, T. Liu, W. Pan, B. Hu, and T. Zhu, "Acoustic differences between healthy and depressed

people: a cross-situation study," *BMC Psychiatry*, vol. 19, no. 1, p. 300, Dec. 2019.
[9] S. Kemper and S. Kemper, "Elderspeak : Speech accommodations to older adults Elderspeak : Speech Accommodations to Older Adults," vol. 5585, 2007.
[10] C. Kitamura, C. Thanavishuth, D. Burnham, and S. Luksaneeyanawin, "Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language," *Infant Behav Dev*, vol. 24, no. 4, pp. 372–392, 2001, doi: 10.1016/S0163-6383(02)00086-3.
[11] E. Lattey, "Interlinguistic Variation and Similarity in Foreigner Talk," in *The Dynamic Interlanguage*, Boston, MA: Springer US, 1989, pp. 87–100.
[12] M. Garnier, N. Henrich, and D. Dubois, "Influence of Sound Immersion and Communicative Interaction on the Lombard Effect," *Journal of Speech, Language, and Hearing Research*, v 53, no. 3, pp. 588–608, 2010
[13] N. Lavan and R. Holloway, "Flexible voices : identity perception from variable vocal signals," *Psychonomic Bulletin & Review*, pp. 1–27, 2013.
[14] S. Cook and J. Wilding, "Earwitness Testimony: Never Mind the Variety, Hear the Length," *Appl Cogn Psychol*, vol. 11, no. 2, pp. 95–111, Apr. 1997.
[15] R. Roebuck and J. Wilding, "Effects of vowel variety and sample lenght on identification of a speaker in a line-up," *Appl Cogn Psychol*, vol. 7, no. 6, pp. 475–481, Nov. 1993, doi: 10.1002/acp.2350070603.
[16] G. E. Legge, C. Grosmann, and C. M. Pieper, "Learning unfamiliar voices.," *J Exp Psychol Learn Mem Cogn*, vol. 10, no. 2, pp. 298–303, Apr. 1984, doi: 10.1037/0278-7393.10.2.298.
[17] P. D. Bricker and S. Pruzansky, "Effects of Stimulus Content and Duration on Talker Identification," *J Acoust Soc Am*, vol. 40, no. 6, pp. 1441–1449, Dec. 1966, doi: 10.1121/1.1910246.
[18] M. Wester, "Talker discrimination across languages," *Speech Commun*, vol. 54, no. 6, pp. 781–790, 2012, doi: 10.1016/j.specom.2012.01.006.
[19] J. M. Zarate, X. Tian, K. J. P. Woods, and D. Poeppel, "Multiple levels of linguistic and paralinguistic features contribute to voice recognition.," *Sci Rep*, vol. 5, p. 11475, 2015, doi: 10.1038/srep11475.
[20] N. Lavan, B. Short, A. Wilding, and C. McGettigan, "Impoverished encoding of speaker identity in spontaneous laughter," *Evolution and Human Behavior*, vol. 39, no. 1, pp. 139–145, Jan. 2018, doi: 10.1016/j.evolhumbehav.2017.11.002.
[21] Z. F. Peynircioğlu, B. E. Rabinovitz, and J. Repice, "Matching Speaking to Singing Voices and the Influence of Content," *Journal of Voice*, vol. 31, no. 2, pp. 256.e13-256.e17, Mar. 2017, doi: 10.1016/j.jvoice.2016.06.004.
[22] C. R. Narayan, L. Mak, and E. Bialystok, "Words Get in the Way: Linguistic Effects on Talker Discrimination," *Cogn Sci*, vol. 41, no. 5, pp. 1361–1376, Jul. 2017, doi: 10.1111/cogs.12396.
[23] 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.
[24] A. R. Reich and J. E. Duke, "Effects of selected vocal disguises upon speaker identification by listening," *J Acoust Soc Am*, vol. 66, no. 4, pp. 1023–1028, 1979.
[25] A. Hirson and M. Duckworth, "Glottal fry and voice disguise: a case study in forensic phonetics," *J Biomed Eng*, vol. 15, no. 3, pp. 193–200, May 1993.
[26] A. Bartle and V. Dellwo, "Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech," *International Journal of Speech, Language and the Law*, vol. 22, no. 2, pp. 229–248, 2015.
[27] H. M. J. Smith, T. S. Baguley, J. Robson, A. K. Dunn, and P. C. Stacey, "Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance," *Appl Cogn Psychol*, vol. 33, no. 2, pp. 272–287, Mar. 2019, doi: 10.1002/acp.3478.
[28] A. Afshan, J. Kreiman, and A. Alwan, "Speaker discrimination performance for 'easy' versus 'hard' voices in style-matched and -mismatched speech," *J Acoust Soc Am*, vol. 151, no. 2, pp. 1393–1403, Feb. 2022, doi: 10.1121/10.0009585.
[29] A. Afshan, J. Kreiman, and Alwan Abeer, "Speaker discrimination in humans and machines: Effects of speaking style variability," in *INTERSPEECH 2020*, Shangai, China, 2020.
[30] S. v. Stevenage, R. Tomlin, G. J. Neil, and A. E. Symons, "May I Speak Freely? The Difficulty in Vocal Identity Processing Across Free and Scripted Speech," *J Nonverbal Behav*, vol. 45, no. 1, pp. 149–163, 2021
[31] S. J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P. A. Keating, and A. Alwan, "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," *J Acoust Soc Am*, vol. 144, no. 1, pp. 375–386, Jul. 2018, doi: 10.1121/1.5045323.
[32] T. Kathiresan *et al.*, "Mothers Reveal More of Their Vocal Identity When Talking to Infants," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4088888.
[33] L. Lansade, M. Trösch, C. Parias, A. Blanchard, E. Gorosurreta, and L. Calandreau, "Horses are sensitive to baby talk: pet-directed speech facilitates communication with humans in a pointing task and during grooming," *Anim Cogn*, vol. 24, no. 5, pp. 999–1006, Sep. 2021, doi: 10.1007/s10071-021-01487-3.
[34] J. B. Grier, "Nonparametric indexes for sensitivity and bias: Computing formulas.," *Psychol Bull*, vol. 75, no. 6, pp. 424–429, 1971, doi: 10.1037/h0031246.
[35] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J Stat Softw*, vol. 67, no. 1, 2015.
[36] G. P. M. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," *Speech Commun*, vol. 22, no. 1, pp. 43–65, Jul. 1997, doi: 10.1016/S0167-6393(97)00012-5.