Title

Speaker discrimination performance for "easy" versus "hard" voices in style-matched and -mismatched speech.

Journal

The Journal of the Acoustical Society of America, 151(2)

ISSN

0001-4966

Authors

Afshan, Amber
Kreiman, Jody
Alwan, Abeer

Publication Date

2022-02-01

DOI

10.1121/10.0009585

Peer reviewed

FEBRUARY 28 2022

Speaker discrimination performance for “easy” versus “hard” voices in style-matched and -mismatched speech

Amber Afshan; Jody Kreiman; Abeer Alwan

J Acoust Soc Am 151, 1393–1403 (2022)
https://doi.org/10.1121/10.0009585
CHORUS

JASA ARTICLE

Speaker discrimination performance for "easy" versus "hard" voices in style-matched and -mismatched speech

Amber Afshan,[1,a)] Jody Kreiman,[2,b)] and Abeer Alwan[1]
[1]Department of Electrical and Computer Engineering, University of California, Los Angeles, California 90095-1594, USA
[2]Departments of Head and Neck Surgery and Linguistics, University of California, Los Angeles, California 90095-1794, USA

ABSTRACT:
This study compares human speaker discrimination performance for read speech versus casual conversations and explores differences between unfamiliar voices that are "easy" versus "hard" to "tell together" versus "tell apart." Thirty listeners were asked whether pairs of short style-matched or -mismatched, text-independent utterances represented the same or different speakers. Listeners performed better when stimuli were style-matched, particularly in read speech−read speech trials (equal error rate, EER, of 6.96% versus 15.12% in conversation–conversation trials). In contrast, the EER was 20.68% for the style-mismatched condition. When styles were matched, listeners' confidence was higher when speakers were the same versus different; however, style variation caused decreases in listeners' confidence for the "same speaker" trials, suggesting a higher dependency of this task on within-speaker variability. The speakers who were "easy" or "hard" to "tell together" were not the same as those who were "easy" or "hard" to "tell apart." Analysis of speaker acoustic spaces suggested that the difference observed in human approaches to "same speaker" and "different speaker" tasks depends primarily on listeners' different perceptual strategies when dealing with within- versus between-speaker acoustic variability. © *2022 Acoustical Society of America*.
https://doi.org/10.1121/10.0009585
(Received 22 July 2021; revised 22 January 2022; accepted 28 January 2022; published online 28 February 2022)
[Editor: Melissa Michaud Baese-Berk] Pages: 1393–1403

I. INTRODUCTION

The manner in which a speaker says an utterance can change unintentionally from one scenario to another, for example, due to social context (e.g., talking to a friend versus public speaking) or emotional or physiological state; or it can change intentionally, for example, to express irony or in an attempt to hide one's identity (Kreiman and Sidtis, 2011). These variations introduce within-speaker variability that strongly impacts perception of unfamiliar voices (Lavan *et al.*, 2019a; Lavan *et al.*, 2019b). The effects of speaking style variability on recognition accuracy have been studied extensively, particularly in the forensic literature (Blatchford and Foulkes, 2006; González Hautamäki *et al.*, 2019; González Hautamäki *et al.*, 2015; Saslove and Yarmey, 1980). For example, style variability confuses ear witnesses hearing a suspect shouting versus reading aloud during a voice lineup (Jessen, 2008). In non-forensic work, humans consistently outperformed machines in both style-matched and -mismatched conditions when discriminating speakers from samples of read versus pet-directed speech (characterized by exaggerated prosody; Park *et al.*, 2018), although style variations resulted in worse performance in both humans and machines. Differences in style were extreme in both these examples. However, knowledge about the effects of moderate variations in speaking style (e.g., between read and conversational speech) on human speaker recognition and discrimination performance is limited. In this paper, we examine the effects of such variations in speaking style on human speaker discrimination performance for unfamiliar voices, using short duration (∼3 s), text-independent utterances.

Two recent studies have provided insights into the ways in which listeners deal with moderate speaker variability. Smith *et al.* (2019) compared style-matched read speech trials with read versus spontaneous speech trials and found that listeners were more accurate and confident in style-matched trials compared to style-mismatched ones. However, their experiments included style-matched trials only from read speech, leaving open the question of which kind of speech better allows listeners to extract a single identity. A second study (Stevenage *et al.*, 2021) addressed this limitation by including style-matched spontaneous speech as well. They found that performance on style-matched trials exceeded that for mismatched trials, with performance on style-matched read speech trials better than that for style-matched spontaneous speech. Their results also revealed a significant bias toward "same speaker" over "different speaker" responses.

Although these studies show that acoustic variability confuses listeners, they leave open the important questions of why and how this occurs. Neither paper quantified the extent of acoustic variability between the two types of trials, nor did they examine the relationship between acoustic

[a)]Electronic mail: amberafshan@g.ucla.edu, ORCID: 0000-0002-1560-8559.
[b)]ORCID: 0000-0002-5360-1729.

variability and how well listeners performed in "same speaker" versus "different speaker" tasks.

Evidence from voice sorting tasks indicates that humans do vary their perceptual strategies when "telling people together" (i.e., assessing within-speaker variability in voice) versus "telling people apart" (i.e., assessing between-speaker variability in voice) (Johnson *et al.*, 2020; Lavan *et al.*, 2019b). However, we do not know how or why listeners vary their perceptual strategies in trials where speakers are the same (a "same speaker" task) versus trials in which speakers are different (a "different speaker" task). In "same speaker" trials, differences between stimuli reflect within-talker acoustic variability, while in "different speaker" trials differences largely reflect between-speaker variability. The nature and extent of differences in listener performance in these two trial types should follow from differences in the nature and extent of these two kinds of variability. Thus, three major questions arise: (i) How does human speaker discrimination performance vary with speaking style? (ii) Is there a difference in how speaking style variations affect trials where speakers are the same versus different? (iii) How does human speaker perception relate to the nature and extent of acoustic variability that occurs within- versus between-speakers?

Recent studies (Lee *et al.*, 2019; Lee and Kreiman, 2019) showed that the most important principal components describing acoustic variability for individual speakers were shared by all the speakers, but the majority of the principal components were idiosyncratic. Moreover, individual speakers' acoustic spaces (within-speaker variability) and spaces for whole populations of speakers (between-speaker variability) shared a similar structure. This shared structure was mainly computed over the balance of higher-frequency harmonic versus inharmonic energy in the voice and over formant dispersion in read speech. In conversational speech, the structure corresponded to variability in source spectral shape, in spectral noise, in $F0$, and in higher formant frequencies. However, little is known about the relationships among within- and between-speaker acoustic variability and listener performance, particularly in the context of differences in speaking style. In this study, we examined these relationships by asking listeners to discriminate among speakers with moderate speaking style variations. Listener performance for individual speakers was interpreted with respect to the speakers' acoustic spaces, with separate analyses for "same speaker" and "different speaker" trials. We hypothesized that speaking style variability would have a large effect on performance in the case of unfamiliar speaker discrimination because the "same speaker" task largely relies on within-speaker variability. Moreover, casual conversations have a higher degree of variation in comparison to read speech (Lavan *et al.*, 2019c), suggesting that the "same speaker" task may be more difficult for conversational speech. Performance on "different speaker" tasks theoretically relies on the relative positions of voices in a shared acoustic structure (between-speaker variability). Previous research (Laan, 1992) has shown that there are inconsistencies between listeners when classifying read and conversational speech, indicating that the moderate differences between these styles have minor perceptual effects, and suggesting that moderate speaking style variations result in small within-speaker variability. Based on this work and on the studies just reviewed, we hypothesized that moderate speaking style variability would have a smaller effect on speaker discrimination performance for "different speaker" trials as they primarily rely on between-speaker variability.

This study extends our previous work (Afshan *et al.*, 2020) and the studies by Stevenage *et al.* (2021) and Smith *et al.* (2019), which compared human and machine speaker discrimination. Using the perceptual data reported in more detail here, Afshan *et al.* (2020) compared human and machine results both at the system-level and speaker-level for style-matched and -mismatched trials, while the present study focuses entirely on the factors governing human voice discrimination.

II. METHODS

A. Perceptual speaker discrimination

1. Stimuli

Voice samples from 40 female speakers (also used in Park *et al.*, 2018; Park *et al.*, 2019) were drawn from the UCLA Speaker Variability Database (Keating *et al.*, 2019; Keating *et al.*, 2021; Kreiman *et al.*, 2015), which incorporates commonly-occurring variations in voice deriving from phonetic content, speaking style, and affect conditions. This database includes speech from 101 female and 101 male speakers, recorded with a 1/2 in. Brüel & Kjær microphone in a sound-attenuated booth at a sampling rate of 22 kHz. Forty speakers were studied to balance concerns about testing duration versus sampling considerations and to provide continuity with our previous perception experiments using this dataset (Park *et al.*, 2018; Park *et al.*, 2019). Samples were restricted to female speakers to avoid any gender-dependent cues, and because females produced clearer contrasts between speaking styles than male speakers did (as judged by the authors). All speakers were self-reported native speakers of American English (confirmed *post hoc* by two linguists). Two sets of voice samples were selected for each speaker. The first (clear read speech) included five phonetically-rich Harvard sentences (IEEE Subcommittee on Subjective Measurements, 1969), read twice in random order. The second (casual conversational speech) consisted of the speakers' side of a 2-min telephone conversation with a family member or friend. The recordings were post-processed to remove any long preceding or trailing silences and all non-speech vocalizations (laughing, giggling, sighing). Six ∼3 s clips were taken from each recording. Selections were carefully made to ensure that semantic cues would not bias responses. For instance, stimuli were chosen from different topics in the conversation. All chosen stimuli were recorded on the same day.

https://doi.org/10.1121/10.0009585

2. Listeners and listening task

All experimental procedures were approved by the UCLA Institutional Review Board. Thirty normal-hearing listeners including 24 native speakers of English (22 female, 8 male) participated in this experiment.[1] The sample size was determined such that there are 12–15 listeners per set of voices.

Each listener undertook three kinds of comparisons, in random order. In one they heard two different read sentences; in another, they compared two different clips excerpted from a conversation; and in the third, they compared one read sentence and one conversational sentence. Equal numbers of "same speaker" and "different speaker" trials were included for each of these three trial types, resulting in six different kinds of trials per experiment. Care was taken to make sure that a listener never heard the same stimulus twice. As only five different sentences had been recorded in the case of read speech, we randomly chose a second recording of one of the five sentences to repeat for the sixth trial.

Listeners were tested individually in a sound-attenuated booth. Stimulus pairs were played in random order over Etymotic insert earphones (model ER-1) at a constant comfortable listening level. To minimize fatigue, listeners heard one of two subsets of speakers (15 listeners per subset). Each subset included 24 speakers selected at random from the pool of 40, for a total of 144 trials per listener (6 trial types × 24 speakers). On "different speaker" trials, two speakers were paired at random, such that each was compared with every other speaker an equal number of times. Each listener heard the stimulus pairs in a unique random order and was asked (i) "Did the two voices represent the same speaker or two different speakers?, and (ii) "How confident are you in your response on a scale of 0 to 5 (0 = wild guess and 5 = very confident)?" Pairs of stimuli could be heard twice, once in each presentation order (AB/BA). Listeners were not aware of the number of speakers included in the experiment. They were encouraged to complete the experiments at their own pace, taking breaks as necessary. Testing time averaged about 45 min.

B. Evaluation metric

1. Calculation of scores

Same/different responses were combined with confidence ratings to create an unfolded similarity score for each stimulus pair. Confidence ratings (0 to 5) were multiplied by the decision (different = −1 and the same = 1) to provide continuous scores ranging from −5 (highly confident that the voices are different) to 5 (highly confident that the voices are the same). This ensured that the similarity score reflected listeners' confidence as well as their same/different decisions.

Similarity scores were used to calculate calibrated log-likelihood ratios (LLRs), denoted as L. LLRs are used in this work instead of similarity scores by themselves, as they provide reliable probabilistic interpretations of the comparisons of the two hypotheses ("same speaker" or "different speaker"). Thus, LLRs provide a single identification score that can be meaningfully interpreted. In other words, calibrated LLRs provide numerical representations of listeners' degree of support for either hypothesis in each trial. This allows us to measure not only their discriminating power but also the strength of the trials evaluated by them (Ramos *et al.*, 2011). Moreover, the calibrated LLRs are needed to obtain the log-likelihood-ratio cost function in Sec. II B 2, which unlike the standard measures is application-independent. This provides a universal probabilistic interpretation in the analysis. A calibration system based on a standard logistic regression solution (Brümmer and De Villiers, 2011a) was used to estimate the LLRs by optimizing the following mapping:

$$L_t = a + bs_t, \tag{1}$$

where L_t is the calibrated output log-likelihood-ratio for trial t and s_t is the similarity score for trial t. Offset parameter a and the weight b are optimized with logistic regression (Brümmer, 2010).

2. Analysis of performance errors

Speaker discrimination performance was evaluated in terms of equal error rates (EER) and the log-likelihood-ratio cost function (C_{llr}) (Van Leeuwen and Brümmer, 2007). While the EER is a widely-used measure, it does not measure ability to set good decision thresholds. Hence, C_{llr}, an application-independent measure for evaluating soft decisions, was also used. It can be interpreted as a measure of insufficiency of information: The lower the C_{llr}, the more the average information per trial (in bits) increases. In Van Leeuwen and Brümmer (2007), a closed-form solution for C_{llr} is provided,

$$C_{\text{llr}}(L_t) = \frac{1}{2}\left(\sum_{t\in\text{same}} \frac{\log_2(1+e^{-L_t})}{N_{\text{same}}} + \sum_{t\in\text{diff}} \frac{\log_2(1+e^{L_t})}{N_{\text{diff}}}\right), \tag{2}$$

where L_t is the log-likelihood-ratio for trial t, "same" is a set of N_{same} "same speaker" trials and "diff" is a set of N_{diff} "different speaker" trials. These two normalized terms represent the costs for "same speaker" (first term) and "different speaker" (second term) trials.

We used the Bosaris toolkit (Brümmer and De Villiers, 2011b) to perform calibration and to calculate the evaluation measures. As data were limited, the calibration parameters were trained on and applied to the same set of scores.

3. Speaker-level analysis

This section describes speaker-level equivalents of the measures described in the previous sections. The LLR, L_t, which represents listeners' scalar responses to each given trial, was obtained for each trial t, as outlined in Sec. II B 1. To compare the scores for "same speaker" and "different

speaker" trials involving each speaker, L^{same} for "same speaker," and L^{diff} for "different speaker" trials were calculated separately. L^{same} for a speaker was obtained by averaging the L_t values over the "same speaker" trials that included that particular speaker. It measures within-speaker variability across the stimuli as perceived by the listeners: a large L^{same} means small perceived within-speaker variability (i.e., these "same speaker" trials are easy). L^{diff} for a given speaker was calculated by averaging the L_t values over the "different speaker" trials that included a given speaker; it represents between-speaker variability across the stimuli as perceived by the listeners. A large L^{diff} value indicates that the speaker has small perceived between-speaker variability, making it difficult for listeners to distinguish her from others.

A speaker-level aggregation of the log-likelihood-ratio cost function (C_{llr}; Sec. II B 2) was also computed by calculating the mean across listeners over all the trials that included that particular speaker. The speaker-level C_{llr} represents the confidence listeners had when identifying that speaker. Speaker-level C_{llr} values for "same speaker" trials ($C_{\text{llr}}^{\text{same}}$) and "different speaker" trials ($C_{\text{llr}}^{\text{diff}}$) were also computed by calculating the average of their respective C_{llr} values across listeners.

For the speaker-level analysis, the trials were combined across conditions due to the limited number of trials per speaker (nine trials × number of listeners). Although collapsing conditions in this way precludes examination of factors other than speaker, these analyses focus primarily on the main effect of differences among speakers, so we felt adding power to tests of this main effect outweighed other considerations. Note that the system-level values are denoted by using a ($'$), i.e., C'_{llr}, $C_{\text{llr}}^{\prime\text{same}}$, $C_{\text{llr}}^{\prime\text{diff}}$, L', $L^{\prime\text{same}}$, and $L^{\prime\text{diff}}$.

C. Speaker acoustic variability

1. Feature extraction and data processing

Acoustic measures were extracted for 25 ms frames, with a 5 ms frame-shift, from utterances of vowels and approximants (i.e., /l/, /r/, /w/) in the stimuli using VoiceSauce (Shue *et al.*, 2011). Feature selection was motivated by a psychoacoustic model of voice quality (Garellek *et al.*, 2016; Kreiman *et al.*, 2021). The set comprised fundamental frequency (F_0), the first four formants (F_1, F_2, F_3, F_4), cepstral peak prominence (*CPP*; Hillenbrand *et al.*, 1994), and the amplitude differences between the first (H_1^*), second (H_2^*), and fourth (H_4^*) harmonics, and the harmonics nearest 2 kHz (H_{2k}^*) and 5 kHz (H_{5k}), denoted as $H_1^* - H_2^*$, $H_2^* - H_4^*$, $H_4^* - H_{2k}^*$, and $H_{2k}^* - H_{5k}$. These measures quantified the harmonic source spectral shape. Harmonic values marked with * were corrected for the influence of formants on harmonic amplitudes (Hanson and Chuang, 1999; Iseli and Alwan, 2004). Following Lee *et al.* (2019) and Lee and Kreiman (2019), we also included formant dispersion (*FD*, calculated as the average difference in the frequency between each adjacent pair of formants), *energy* (a measure of amplitude given by root mean square energy calculated over five pitch pulses), and the ratio of amplitudes of subharmonics to harmonics (*SHR*; Sun, 2002; Herbst, 2021) as a measure of period doubling, for a total of 13 features for every analysis frame.

Frames with missing or unrealistic values were removed, after which features were normalized with reference to global maxima and minima, for a range across speakers of 0–1. We then calculated the moving average and the moving coefficient of variation (moving CoV = moving standard deviation/moving average) over a 25 ms window (commonly used in speech feature extraction, equivalent to five observations) for each of the 13 features. This resulted in a total of 26 acoustic features (13 moving averages and 13 moving CoVs). These 26 features were used for subsequent analysis.

2. Principal component analysis

Figure 1 represents the block diagram of the speaker variability analysis. Following Lee *et al.* (2019) and Lee and Kreiman (2019), we applied principal component analysis (PCA) to characterize acoustic variability in the voices of individual speakers. Utterances from each speaker were used to calculate the within-speaker PCA representing that individual's acoustic space. We retained only the principal components with eigenvalues greater than one so that each represented an interpretable amount of variance in the data (Kaiser, 1960).

The analytical approach proposed by Krzanowski (1979) was used to compare PCA spaces, to avoid reliance on subjective criteria associated with visual examination. In this approach, let g be the number of speakers being compared with n_t observations for the tth speaker ($t = 1, 2, \ldots, g$), with the same set of p variables measured for each speaker. Let us assume that for each speaker, k_t principal components represent that speaker's acoustic variability. Next, let b be an arbitrary vector in the original p-dimensional data-space and let δ_t be the angle between b and the vector most parallel to it in the space generated by the k_t principal components of speaker t ($t = 1, 2, \ldots, g$). We represent the loadings using the matrix L_t where the element $l_{ij}^{(t)}$ represents the loading of the jth variable on the ith principal component of the tth speaker. Then the value of b that minimizes $V = \sum_{t=1}^{g} \cos \delta_t{}^2$ is given by the eigenvector b_1, corresponding to the largest eigenvalue μ_1 of $H = \sum_{t=1}^{g} L_t' L_t$.

The eigenvector b_2, corresponding to the second largest eigenvalue of H, satisfies the criterion for the next largest value of V and is orthogonal to b_1. When k_t different components have been obtained for the tth speaker ($t = 1, 2, \ldots, g$) and $k = \min(k_1, k_2, \ldots, k_g)$, then only a k-dimensional comparison will be useful. Any further dimension will be orthogonal to at least one of the speaker spaces. Using this transformation thus allows us to compare different principal component subspaces, because the eigenvalues μ_i [alternatively, the minimum angles $\cos^{-1}(\mu_i)^{1/2}$] can provide a measure of the extent to which the subspaces differ, and the

https://doi.org/10.1121/10.0009585

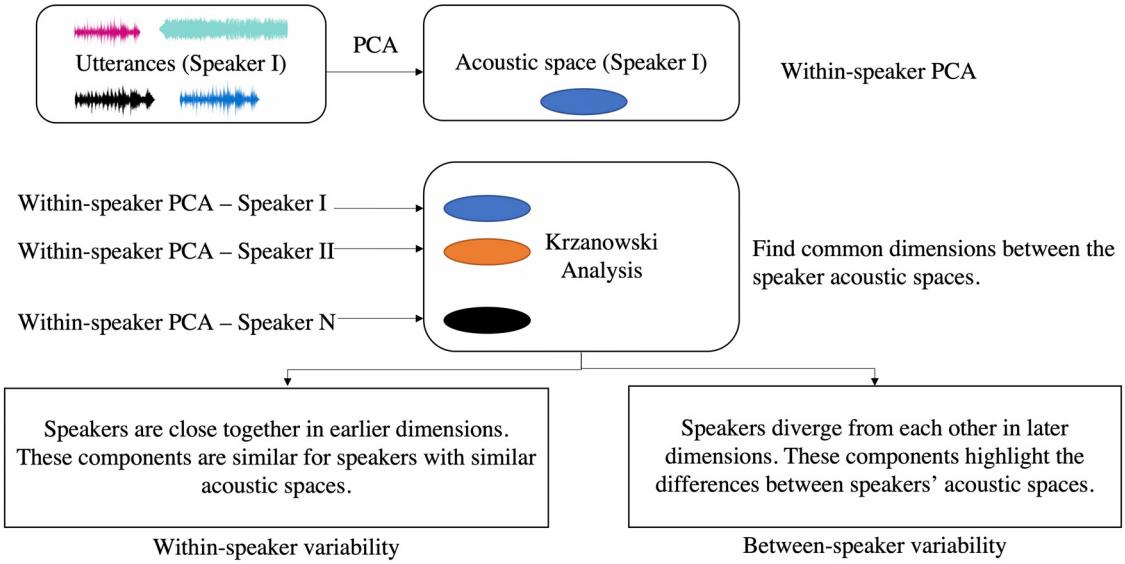

FIG. 1. (Color online) Block diagram representing the analysis of variability in speaker acoustic spaces using PCA and Krzanowski analysis.

eigenvectors b_i can describe the nature of their similarities or differences. The smaller the angles between the subspaces, the higher the similarity. The Appendix provides an overview of the Krzanowski analysis implementation for one set of speakers.

Krzanowski analysis was performed over the within-speaker PCAs for all the speakers in a set, to obtain the dimensions common to the different speaker acoustic spaces. The earlier (lower) dimensions represent the components that are similar for speakers with similar acoustic spaces and which are thus suitable for comparing within-speaker variability across speakers. The speakers diverge from each other in later (higher) dimensions. Hence, the components in these dimensions highlight the differences between speakers' acoustic spaces, making them appropriate indices of between-speaker variability.

III. RESULTS

Table I shows speaker discrimination performance for the three speaking-style conditions (read speech–read speech, conversation–conversation, and read speech–conversation). Statistical significance was evaluated using a two-sample Kolmogorov-Smirnov test (Smirnov, 1948). All reported comparisons are statistically significant at the $p < 0.05$ level. EER values in this table indicate that listeners performed best when voice samples were style-matched read speech (EER = 6.96%). Performance decreased for conversation–conversation trials (EER = 15.12%; $p = 0.035$, $D = 0.059, N = 2304$), even though these were also style-matched. This decrease in performance is likely due to additional variability in casual conversations (formal/informal, happy/sad/angry/neutral, etc.; Lavan *et al.*, 2019c). The style-mismatched read speech − conversation trials resulted in performance that was significantly worse than in either style-matched condition (read speech: $p = 4.95 \times 10^{-14}$, $D = 0.164, N = 2304$; conversation: $p = 4.61 \times 10^{-7}$, $D = 0.115, N = 2304$).

A comparison of the log-likelihood-ratio cost functions (see Sec. II B 2), $C'^{\text{same}}_{\text{llr}}$, and $C'^{\text{diff}}_{\text{llr}}$ values in Table I indicates that "same speaker" trials were easier than "different speaker" trials in all conditions (read speech–read speech: $p = 1.4 \times 10^{-101}, D = 0.88, N = 1152$; conversation–conversation: $p = 4.76 \times 10^{-73}, D = 0.72, N = 1152$; read speech–conversation: $p = 9.07 \times 10^{-74}, D = 0.59, N = 1152$). Differences in difficulty between the two tasks depended on speaking style, with style-matched read speech − read speech trials showing the best performance overall (0.210 and 0.318 for $C'^{\text{same}}_{\text{llr}}$ and $C'^{\text{diff}}_{\text{llr}}$, respectively) and the most difference between the same and different speaker tasks.

A. Speaker-level log-likelihood-ratio analysis

Figure 2 compares the distribution kernel density plots overlaid onto histograms of speaker-level log-likelihood-ratios (see Sec. II B 3) for "same speaker" (L^{same}) and "different speaker" (L^{diff}) trials for the three style conditions. Recall that the positive end of this scale represents

TABLE I. Speaker discrimination performance in terms of EER (%) and LLR cost function for combined (C'_{llr}), "same speaker" trials ($C'^{\text{same}}_{\text{llr}}$), and "different speaker" trials ($C'^{\text{diff}}_{\text{llr}}$). The better (lower cost) value for "same speaker" versus "different speaker" trials in each condition is underlined. All reported comparisons are statistically significant at the $p < 0.05$ level.

read–read				conversation–conversation				read–conversation			
EER %	C'_{llr}	$C'^{\text{same}}_{\text{llr}}$	$C'^{\text{diff}}_{\text{llr}}$	EER %	C'_{llr}	$C'^{\text{same}}_{\text{llr}}$	$C'^{\text{diff}}_{\text{llr}}$	EER %	C'_{llr}	$C'^{\text{same}}_{\text{llr}}$	$C'^{\text{diff}}_{\text{llr}}$
6.96	0.264	<u>0.210</u>	0.318	15.12	0.529	<u>0.501</u>	0.557	20.68	0.691	<u>0.690</u>	0.692

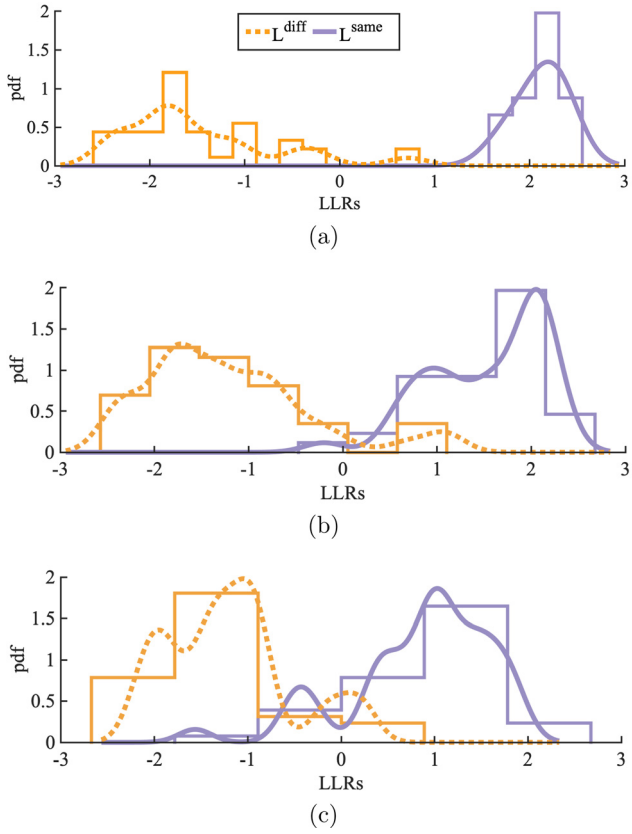

FIG. 2. (Color online) Distribution kernel density plots overlaid onto histograms of speaker-level LLRs for "same speaker" (L^{same}) and "different speaker" (L^{diff}) trials represented as probability density functions. L^{same} and L^{diff} are denoted with solid ("–") and dotted (".") lines, respectively. (a) read speech–read speech; (b) conversational speech–conversational speech; (c) read speech–conversational speech.

highly confident "same" responses, and the negative end represents highly confident "different" responses. The means of L^{same} and L^{diff} are shifted towards correct responses in the read speech − read speech conditions. This increased separation of "same speaker" and "different speaker" trial distributions indicates that discrimination was easier and resulted in better performance in the read speech–read speech condition compared to the other two conditions. For example, compare discrimination performance from the distributions in Figs. 2(a) (read speech–read speech) and 2(c) (read speech–conversational speech) (EERs = 6.96% versus 20.68%, respectively). The read speech–read speech trials resulted in an L^{same} distribution with small variance (*variance* = 0.05) confined to the positive response region. This was not the case with L^{diff} (*variance* = 0.57), indicating that listeners were more confident when classifying "same speaker" pairs than "different speaker" pairs.

In comparison, in the conversation − conversation condition [Fig. 2(b)] variance in the L^{same} distribution increased (*variance* = 0.34), and this distribution overlapped with the L^{diff} distribution (*variance* = 0.70), suggesting that listeners' confidence decreased overall with a change in style from read to conversational speech. Finally, in the read speech − conversation condition [Fig. 2(c)] the variance in the L^{same} distribution increased further (*variance* = 0.75), while it decreased slightly in the L^{diff} distribution (*variance* = 0.55). This overall pattern suggests that style affected the listeners' confidence in "same speaker" tasks, but not in "different speaker" tasks.

Given the multimodal shape of the distributions for conversation and style-mismatched tasks, the findings in terms of variances of LLRs were helpful. We evaluated the differences between the distributions across styles. The speaker-level log-likelihood-ratios for "same speaker" (L^{same}) tasks for the three style conditions differed significantly from one another, with means of 1.8915, 1.4131 and 0.9475 for style-matched read speech, style-matched conversation, and style-mismatched tasks, respectively (read speech–read speech versus conversation–conversation: $h = 1, p = 3.57 \times 10^{-5}$, $D = 0.45, N = 80$, read speech–read speech versus read speech–conversation: $h = 1, p = 7.34 \times 10^{-9}, D = 0.68$, $N = 80$, and conversation–conversation versus read speech–conversation: $h = 1, p = 0.04, D = 0.30, N = 80$). In contrast, the speaker-level LLRs for "different speaker" (L^{diff}) tasks for the three style conditions did not differ significantly (means = −1.4891, −1.3433, and −1.2165 for style-matched read speech, style-matched conversation and style-mismatched tasks, respectively; read speech–read speech versus conversation–conversation: $h = 0, p = 0.14$, $D = 0.25, N = 80$, read speech–read speech versus read speech–conversation: $h = 0, p = 0.08, D = 0.28, N = 80$, and conversation–conversation versus read speech–conversation: $h = 0, p = 0.72, D = 0.15, N = 80$). This result is consistent with our hypothesis that the effect of speaking style-variability is greater in "same speaker" tasks than in "different speaker" tasks.

B. Speaker-level log-likelihood-ratio cost analysis

Recall that the speaker-level log-likelihood-ratio cost function, C_{llr}, denotes the overall speaker information available when the listener is performing speaker discrimination. It is calculated by averaging the values for "same speaker" trials (C_{llr}^{same}) and "different speaker" trials (C_{llr}^{diff}) for a given speaker. A higher C_{llr} indicates less information available to the listener for the speaker discrimination task, hence more difficulty. For "same" and "different" trials, speaker-level C_{llr} values from LLR scores were used to group speakers into three subsets.[2] We classified the 13 speakers with the lowest C_{llr} values ("same speaker" task: mean = 0.251; range = 0.169–0.361, "different-speaker" task: mean = 0.243; range = 0.127–0.367) into an "easy" subset and the 13 speakers with the highest C_{llr} values ("same speaker" task: mean = 0.964; range = 0.669–1.434, "different-speaker" task: mean = 1.076; range = 0.741–1.582) as "hard" (difficult to distinguish speakers). The remaining fourteen speakers were referred to as "average" ("same speaker" task: mean = 0.508; range = 0.368–0.647, "different-speaker" task: mean = 0.538; range = 0.370–0.708).

The joint distribution of speakers across the three subsets for the "same speaker" versus "different speaker" tasks

https://doi.org/10.1121/10.0009585

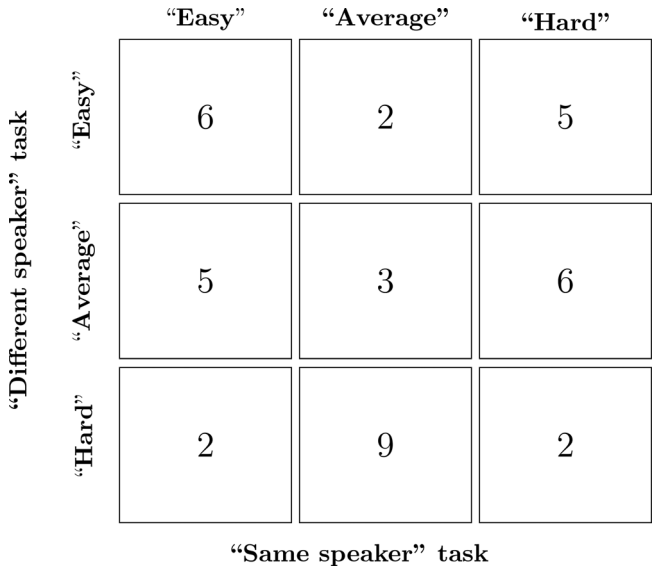

FIG. 3. The number of speakers that were "easy" versus "average" or "hard," as indexed by overall accuracy, for "different speaker" versus "same speaker" tasks. Columns show the number of speakers who were easy, average, or hard to "tell together" on the "same speaker" trials, while rows show how difficult the same voices were to "tell apart" on the "different speaker" trials.

is shown in Fig. 3. An entry $count_{m_i,h_j}$ denotes the number of speakers from subset i of the "different speaker" task overlapping with subset j of the "same speaker" task. For example, in the first column, six samples were "easy" in both the "same speaker" and "different speaker" tasks, whereas two samples that were "easy" in the "same speaker" task were "hard" in the "different speaker" task. More observations falloff diagonal (speakers are not equally "easy" to "tell together" and "tell apart") than on diagonal (the tasks are equally "easy" for that speaker), consistent with findings that humans rely on different information when performing the two tasks (Johnson *et al.*, 2020; Lavan *et al.*, 2019b).

C. Variability in the speaker acoustic spaces

Because the acoustic signal is the input to human perceptual processes, examination of acoustic variability may provide insight into the perceptual strategies listeners use when performing "same speaker" and "different speaker" tasks. To address this, we used PCA to generate principal component subspaces and applied Krzanowski analysis (Sec. II C 2) to compare acoustic variability for speakers who were "easy," "average," or "hard" to discriminate. As noted above, Krzanowski analysis provides a means of quantifying the similarity of the acoustic spaces for different talkers, by generating loadings of the directions in the acoustic spaces that are closest to the PCs for the speakers in each subset.

Figures 4 and 5 show each orthogonal direction as a separate subplot. "Same speaker" trials are shown in Fig. 4 and "different speaker" trials are shown in Fig. 5. The angles listed at the top of each subplot quantify the degree of similarity between all speakers in the set. For ease of comparison, each subplot shows the same dimension for all three subsets of speakers ("easy," "average," and "hard" to discriminate). Speaking styles are combined, however, because speaker C_{llr} values were calculated across all conditions. Plots are additionally restricted to the absolute values of the top three contributing factor loadings to focus attention on the most important contributors to similarity and differences in the acoustic space. Finally, we restricted the number of orthogonal directions to a dimension of $k = 7$, which is the minimum number of principal components extracted per speaker.

1. "Same speaker" task

As Fig. 4 shows, "easy," "average," and "hard" speakers were acoustically similar along the first two dimensions (as indicated by small mean angular separations), but they increasingly diverged after this, with the maximum variation along the 7th dimension. The mean angular separations quantify the extent to which the dimensions represent the similarity between speakers for each subset. Within-speaker variations can be compared along dimension 1, which is associated with CoVs of F_1, F_2, and *FD*. F_2 and *FD* contribute to separating voices on the second dimension; $H_4^* - H_{2k}^*$ also contributed for "easy" speakers, and F_1 for the "average" and "hard" speakers.

Examination of dimension 7 shows that different features underlie acoustic differences for each group of speakers, with mean angular separations of 33.35°, 28.11°, and 29.97° for "easy," "average," and "hard" speakers, respectively. For "easy" speakers, this dimension is related to *Energy*, *Energy* CoV, and F_2 CoV, suggesting that differences between speakers in these factors have little effect on listeners' ability to tell voices together. For "average" speakers, this dimension is related to CoVs of F_0, F_2, and *FD*, while voices that were hardest to tell together varied along F_3, *Energy*, and its CoV. Dimensions 3–6 explained a mixture of similarities and differences, with some speakers closer to each other along those dimensions and others farther apart.

2. "Different speaker" task

Figure 5 compares the principal components describing acoustic variability for speakers classified as "easy," "average," or "hard" to "tell apart" in "different speaker" trials. The coefficients of variation (CoVs) for F_1 and *FD* contributed to separating voices based on their within-speaker variability on the first dimension for all three groups; F_2 CoV also contributed for "easy" and "hard" speakers, while *CPP* CoV contributed for "average" speakers. The second dimension is related primarily to moving averages of F_2 and *FD*. Telling voices apart in the "easy" and "hard" subsets also depended on F_1. Similarity for "average" speakers was also related to $H_4^* - H_{2k}^*$, and "average" speakers were more similar to one another along the second dimension (7.83°) compared to "easy" and "hard" subsets (12.83° and 9.43°, respectively). These results suggest that the means of formant frequencies contribute little to making voices "easy" or "hard" to distinguish in a "different speaker" task.

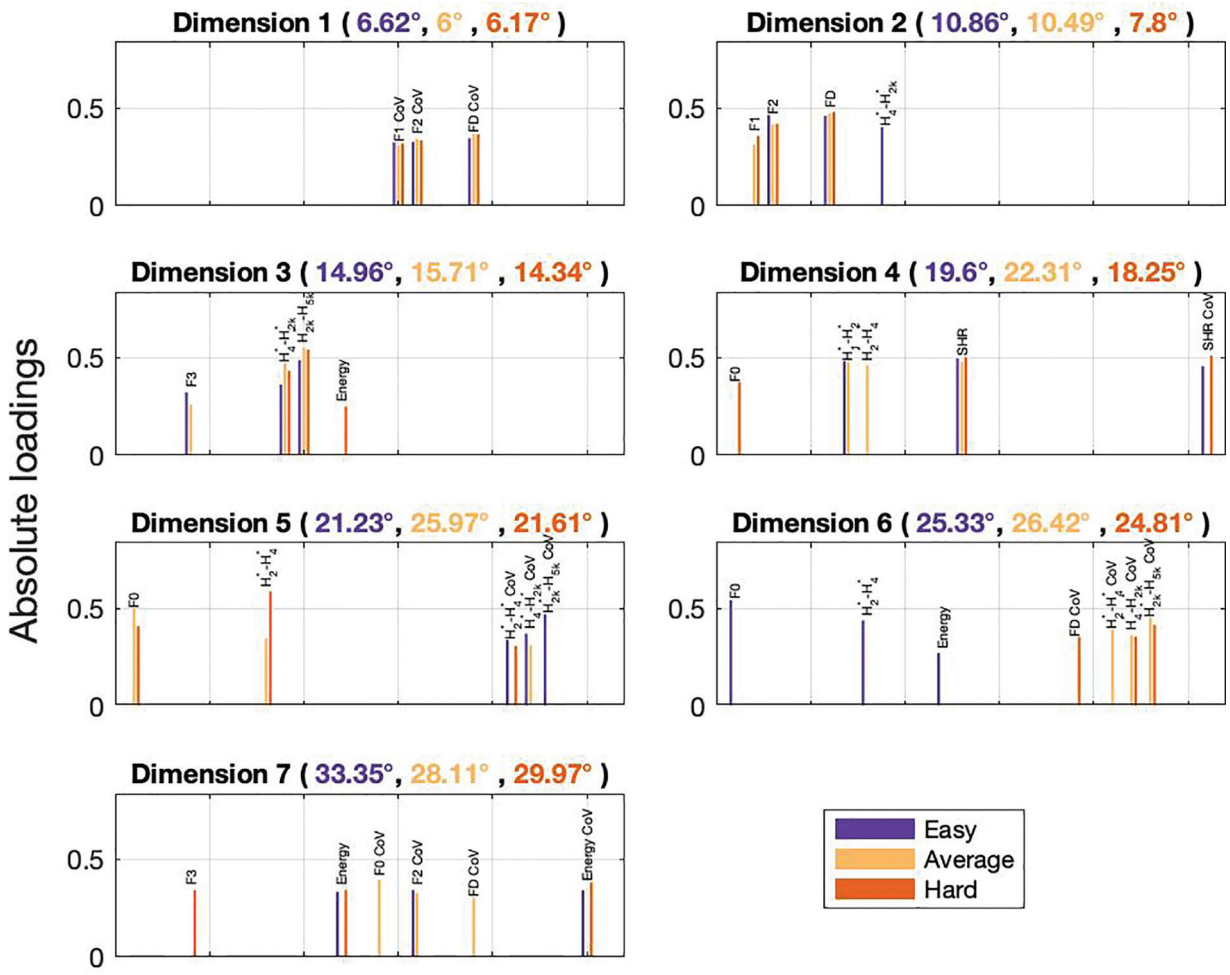

FIG. 4. For the "same speaker" task, the absolute loadings/coefficients of the directions that are closest to the principal components of all speakers in a subset. The mean angular separation between groups and each direction is shown above each subplot. The features are represented along the x axis. F_0, fundamental frequency; F_1, F_2, F_3, F_4, the first four formants; *CPP*, cepstral peak prominence, $H_1^* - H_2^*$, $H_2^* - H_4^*$, $H_4^* - H_{2k}^*$, and $H_{2k}^* - H_{5k}$; the amplitude differences of the harmonics, *FD*, formant dispersion; *SHR*, subharmonics to harmonics ratio; *CoV*, coefficient of variation.

Dimension 7 describes the majority of the between-speaker variability across subsets, with mean angular separations of 28.38°, 31.49°, and 29.75° for "easy," "average," and "hard" speakers, respectively. Speakers who are "easy" to tell apart differed from each other primarily in F_0 CoV, followed by CoVs of $H_4^* - H_{2k}^*$ and $H_{2k}^* - H_{5k}$. In comparison, "average" speakers varied almost equally in terms of F_0, F_2 CoV, and *FD* CoV, while most variation in "hard" speakers was attributable to F_0, followed by smaller contributions from $H_1^* - H_2^*$ and *Energy*. In other words, for the "different speaker" task, discrimination is best for talkers whose speech acoustics are mainly separated by the three CoVs (F_0 CoV, $H_4^* - H_{2k}^*$ CoV and $H_{2k}^* - H_{5k}CoV$), less efficient for "average" talkers whose acoustics differed mainly in mean F_0 and two formant-variable CoVs (extent of variability in relation to the average), and is the worst for talkers whose speech is distinguished only by moving averages.

IV. DISCUSSION

In this paper, we examined the effects of moderate speaking style variations (read speech versus casual conversations) and of within- versus between-speaker acoustic variability on human speaker discrimination performance. The stimuli comprised short text-independent utterances from speakers who were not familiar to the listeners.

The first objective of this work was to identify the effects of speaking style variations on human speaker discrimination performance. Listeners performed better in style-matched cases (EER = 6.96% when both stimuli were read sentences and EER = 15.12% when both stimuli were excerpts from conversations) than in the style-mismatched case (EER = 20.68%). Moderate speaking style variations affected speaker discrimination performance when stimuli were style-mismatched and also when they were style-matched i.e., read speech trials were easier than conversation trials.

In comparison to our previous findings based on read and pet-directed speech from the same speakers (Park *et al.*, 2018) the performance gap between the style-matched and style-mismatched conditions appears to depend at least partly on the extent of the mismatch (moderate in the present study and extreme in the previous study). For example, the EER in Park *et al.* (2018) for the style-matched read

https://doi.org/10.1121/10.0009585

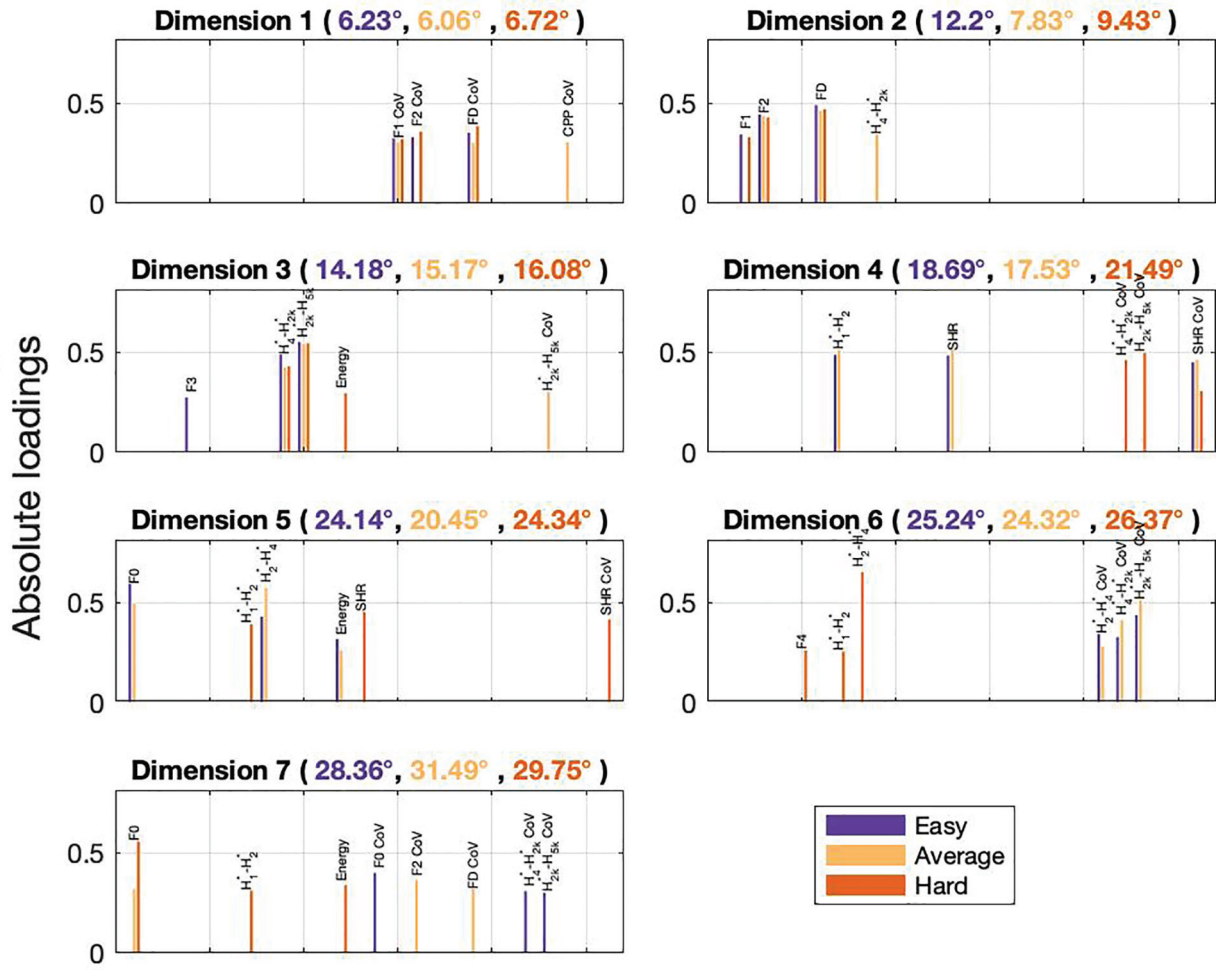

FIG. 5. For the "different speaker" task, the absolute loadings/coefficients of the directions that are closest to the principal components of all speakers in a subset. The mean angular separation between groups and each direction is shown above each subplot. The features are represented along the x axis. F_0, fundamental frequency; F_1, F_2, F_3, F_4, the first four formants; *CPP*, cepstral peak prominence; $H_1^* - H_2^*$, $H_2^* - H_4^*$, $H_4^* - H_{2k}^*$, and $H_{2k}^* - H_{5k}$, the amplitude differences of the harmonics; *FD*, formant dispersion; *SHR*, subharmonics to harmonics ratio; *CoV*, coefficient of variation.

speech–read speech condition was 19.02%, while for the style-mismatched condition it was 39.23%, versus 6.96% and 20.68%, respectively, in the present study.[3]

Another objective of this research was to determine the differences in how speaking style variations affect "same speaker" and "different speaker" trials. The speaker-level LLR distribution (see Fig. 2) skewed heavily toward the positive region with small variance in "same speaker" trials, indicating that listeners were more accurate and more confident in the "same speaker" trials. Confidence on these trials was highest for read speech–read speech and worst for read speech–conversation; this pattern did not occur for "different speaker" trials. The changes in listeners' confidence in "same speaker" trials seem to follow the same pattern as did overall performance. Listeners were highly confident for the style-matched read speech trials, but confidence decreased substantially for the other two conditions. Taken together, these results are consistent with our hypotheses that the "same speaker" task largely relies on within-speaker variability, and that moderate style variations impact human performance. However, no such confidence differences arose from style variability in the "different speaker" trials. This suggests that between-speaker variability in the "different speaker" task has greater influence on human performance compared to the effects of moderate speaking style variability.

We also found that which voices listeners judged most accurately depended not only on the voices but also on the task: "telling speakers together" was easier for some voices, while "telling speakers apart" was easier for others. This suggests that listeners rely on different acoustic information when performing these two tasks. In the "same speaker" task, the "easy" speakers varied widely along F_2 CoV, *Energy*, and its CoV, while "average" speakers varied the most along CoVs of F_0, F_2, and *FD*. Finally, "hard" speakers varied mainly along F_3, *Energy*, and its CoV in this task. The features that made the "same speaker" task easier (F_2 CoV, *Energy*, and its CoV) were the ones that appeared in later dimensions (dimension 3 or higher), i.e., the ones that

contributed to speaker idiosyncrasies in Lee *et al.* (2019) and the acoustic voice space model of Lee and Kreiman (2019). This further suggests that listeners rely on speaker idiosyncrasies for the "same speaker" task. Note that in this task, formant CoVs (F_2 CoV for "easy" speakers and CoVs of F_2 and FD for "average" speakers) played a critical role in assisting listeners in "telling speakers together." Forensic studies (McDougall, 2004) argue that formant frequency variations have relevant speaker identification information as they are determined not only by the shape and size of the vocal tract but also by the speaker's style of configuring articulators for speech.

In the "different speaker" task, "easy" speakers differed in the CoVs of amplitude differences of the higher harmonics ($H_4^* - H_{2k}^*, H_{2k}^* - H_{5k}$) and F_0. These were some of the variability features that described the shared acoustic structure across speakers in Lee *et al.* (2019) and Lee and Kreiman (2019). These results provide further evidence in support of our hypothesis that the distance along the shared acoustic structure is critical for speaker discrimination in the "different speaker" task. Voices that differed in static acoustic properties, including a combination of mean F_0, lower harmonic amplitudes ($H_1^* - H_2^*$), and energy, were difficult for listeners to distinguish. Moreover, average voices were distinguished by both moving average and variability (CoVs) features, implying that variations between speakers along static acoustic properties could be insufficient for listeners to tell them apart, while variations along feature CoVs assisted listeners in this task.[4]

In summary, it seems that listeners find it easier to "tell speakers together" using speaker-specific idiosyncrasies, i.e., we can best explain the performance on the "same speaker" task by the nature and extent of within-speaker variability. In contrast, listeners "tell speakers apart" based on differences in features (alternatively, relative positions) within a shared acoustic structure rather than speaker-specific features. This implies that "telling speakers apart" relies more on the nature and extent of between-speaker variability as the differences here are across acoustic features representing shared variability. Therefore, it should be possible to perform acoustic-based predictions of which voices will be "easy" or "hard" to "tell apart" using the relative positions in the shared acoustic structure. However, similar acoustic-based predictions about "telling together" different samples of a speaker's voice might be challenging, as this would require finding the speaker-specific idiosyncrasies.

One limitation of this work must be noted. The perception experiments used a homogenous panel of listeners (22 female out of 30 listeners with an age range of 17–21 years old). Hence, these findings may not fully generalize to other populations. The results presented nevertheless provide a means of investigating the question of the effects of moderate style-variability on speaker discrimination performance. In the future, a heterogeneous population will be used for the listeners' panel.

V. CONCLUSION

This study examined speaker discrimination performance and the effects of speaking style variations (read speech versus conversational speech) on voice discrimination accuracy. Our results showed that the difficulty of the discrimination task changed with style: the style-matched read speech–read speech condition was easiest, followed by conversation–conversation. The style-mismatched condition resulted in the worst performance. Moderate speaking style variability has more effect on the "same speaker" task than on the "different speaker" task. The same speakers were not "easy" or "hard" to distinguish in the "same speaker" and "different speaker" tasks. Analysis of acoustic variability suggested that the listeners found it easier to "tell speakers together" when they rely on speaker-specific idiosyncrasies and that they "tell speakers apart" based on their relative positions within a shared acoustic space. The study contributes to our understanding of the relationship between human speaker perception and the nature and extent of acoustic variability. The results of this study indicate that we can make acoustic-based predictions of voices that will be "easy" or "hard" to "tell apart," but such predictions cannot be made for "telling speakers together." Hence, further work is needed to model the perception of speaker-specific idiosyncrasies and their relation to speaker identity. A further study could assess how the present results extend in terms of higher variability in speaking styles.

ACKNOWLEDGMENTS

This research was partially supported by NSF Grant No. 1704167 and Grant No. DC01797 from the NIH. Parts of this research were presented at Interspeech, 2020 (Afshan *et al.*, 2020). The authors would like to thank Dr. Yoonjeong Lee for valuable and profound comments.

APPENDIX: KRZANOWSKI ANALYSIS

ALGORITHM 1: Krwazonski analysis for set with g speakers.

$k \leftarrow \min(k_1, \ldots, k_g)$ // k-dimensional comparison
for *speaker t in set* **do**
  $L_t \leftarrow$ normalized loadings of speaker t
  $H \leftarrow H + L_t' L_t$
end
$V \leftarrow$ Eigenvectors(H) /* Loadings of the directions closest to the speakers in the set */
for *variable j in set of p variables* **do**
  $b \leftarrow V_j$ // Eigenvector corresponding to variable j
  for *speaker t in set* **do**
    $c \leftarrow b' * L_t' * L_t * b$
    $\delta_{j,t} \leftarrow \arccos\sqrt{c}$ // Angle between speaker t and direction j
  end
end

https://doi.org/10.1121/10.0009585

[1]An additional six speakers (three were native speakers of Spanish, two of Mandarin, and one of Hindi) were also tested, but were later deleted from the data set because preliminary analyses suggested effects of native language on listener performance. There were not enough data to explore these effects in detail.

[2]The correlation between the speaker-level $C_{\text{llr}}^{\text{same}}$ and $C_{\text{llr}}^{\text{diff}}$ is weak ($r = -0.0892$), hence, our preference for the categorical approach used here, versus treating difficulty as a continuous variable.

[3]Note that the sampling rate was higher in the present study than in our previous work (22 kHz, versus 8 kHz in Park *et al.*, 2018).

[4]In general, the measures characterizing hard-to-distinguish voices are known to be important for speaker characterization [e.g., fundamental frequency and $H_1^* - H_2^*$ correlate with perceived breathiness (Wayland and Jongman, 2003)], but challenges arise in this task given that it involves female-only comparisons. In a female-only comparison, there are smaller variations in F_0 and influence of nasality on $H_1^* - H_2^*$ (Simpson, 2012).

Afshan, A., Kreiman, J., and Alwan, A. (**2020**). "Speaker discrimination in humans and machines: Effects of speaking style variability," in *Proceedings of Interspeech*, October 25–29, Shanghai, China.

Blatchford, H., and Foulkes, P. (**2006**). "Identification of voices in shouting," Int. J. Speech Lang. Law **13**(2), 241–254.

Brümmer, N. (**2010**). "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. thesis, University of Stellenbosch, South Africa.

Brümmer, N., and De Villiers, E. (**2011a**). "The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF," in *Proceedings of the NIST SRE Analysis Workshop*, 6–7 December 2011, Atlanta, GA.

Brümmer, N., and De Villiers, E. (**2011b**). "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," documentation of BOSARIS toolkit 24, https://sites.google.com/site/nikobrummer/ by Agnitio Labs (Last viewed July 24, 2021).

Garellek, M., Samlan, R., Gerratt, B. R., and Kreiman, J. (**2016**). "Modeling the voice source in terms of spectral slopes," J. Acoust. Soc. Am. **139**(3), 1404–1410.

González Hautamäki, R., Hautamäki, V., and Kinnunen, T. (**2019**). "On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise," J. Acoust. Soc. Am. **146**(1), 693–704.

González Hautamäki, R., Kinnunen, T., Hautamäki, V., and Laukkanen, A.-M. (**2015**). "Automatic versus human speaker verification: The case of voice mimicry," Speech Commun. **72**, 13–31.

Hanson, H. M., and Chuang, E. S. (**1999**). "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," J. Acoust. Soc. Am. **106**(2), 1064–1077.

Herbst, C. T. (**2021**). "Performance evaluation of subharmonic-to-harmonic ratio (SHR) computation," J. Voice **35**(3), 365–375.

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (**1994**). "Acoustic correlates of breathy vocal quality," J. Speech Lang. Hear. Res. **37**(4), 769–778.

IEEE Subcommittee on Subjective Measurements (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**(3), 225–246.

Iseli, M., and Alwan, A. (**2004**). "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 17–21, Montreal, Canada, pp. I–669–I–672.

Jessen, M. (**2008**). "Forensic phonetics," Lang. Linguist. Compass **2**(4), 671–711.

Johnson, J., McGettigan, C., and Lavan, N. (**2020**). "Comparing unfamiliar voice and face identity perception using identity sorting tasks," Q. J. Exp. Psychol. **73**(10), 1537–1545.

Kaiser, H. F. (**1960**). "The application of electronic computers to factor analysis," Educat. Psychol. Meas. **20**(1), 141–151.

Keating, P., Kreiman, J., and Alwan, A. (**2019**). "A new speech database for within- and between-speaker variability," in *Proceedings of the ICPhS XIX*, August 5–9, Melbourne, Australia, pp. 736–739.

Keating, P., Kreiman, J., Alwan, A., Chong, A., and Lee, Y. (**2021**). "UCLA speaker variability database," http://www.seas.ucla.edu/spapl/shareware.html\#Data (Last viewed July 20, 2021).

Kreiman, J., Lee, Y., Garellek, M., Samlan, R., and Gerratt, B. R. (**2021**). "Validating a psychoacoustic model of voice quality," J. Acoust. Soc. Am. **149**(1), 457–465.

Kreiman, J., Park, S. J., Keating, P. A., and Alwan, A. (**2015**). "The relationship between acoustic and perceived intraspeaker variability in voice quality," in *Proceedings of Interspeech*, September 6–10, Dresden, Germany, pp. 2357–2360.

Kreiman, J., and Sidtis, D. (**2011**). *Foundations Voice Studies: An Interdisciplinary Approach to Voice Production Perception* (Wiley, New York), pp. 245–246.

Krzanowski, W. J. (**1979**). "Between-groups comparison of principal components," J. Am. Stat. Assoc. **74**(367), 703–707.

Laan, G. P. (**1992**). "Perceptual differences between spontaneous and read aloud speech," Proc. Inst. Phon. Sci. Amsterdam **16**, 65–79.

Lavan, N., Burston, L. F. K., and Garrido, L. (**2019b**). "How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices," Br. J. Psychol. **110**(3), 576–593.

Lavan, N., Burston, L. F. K., Ladwa, P., Merriman, S. E., Knight, S., and McGettigan, C. (**2019a**). "Breaking voice identity perception: Expressive voices are more confusable for listeners," Q. J. Exp. Psychol. **72**(9), 2240–2248.

Lavan, N., Burton, A. M., Scott, S. K., and McGettigan, C. (**2019c**). "Flexible voices: Identity perception from variable vocal signals," Psychonom. Bull. Rev. **26**(1), 90–102.

Lee, Y., Keating, P., and Kreiman, J. (**2019**). "Acoustic voice variation within and between speakers," J. Acoust. Soc. Am. **146**(3), 1568–1579.

Lee, Y., and Kreiman, J. (**2019**). "Within- and between-speaker acoustic variability: Spontaneous versus read speech," J. Acoust. Soc. Am. **146**(4), 3011.

McDougall, K. (**2004**). "Speaker-specific formant dynamics: An experiment on Australian English /aI/," Int. J. Speech Lang. Law **11**(1), 103–130.

Park, S. J., Afshan, A., Kreiman, J., Yeung, G., and Alwan, A. (**2019**). "Target and non-target speaker discrimination by humans and machines," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 12–17, Brighton, UK, pp. 6326–6330.

Park, S. J., Yeung, G., Vesselinova, N., Kreiman, J., Keating, P. A., and Alwan, A. (**2018**). "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," J. Acoust. Soc. Am. **144**(1), 375–386.

Ramos, D., Franco-Pedroso, J., and Gonzalez-Rodriguez, J. (**2011**). "Calibration and weight of the evidence by human listeners. The ATVS-UAM submission to NIST human-aided speaker recognition 2010," in *Proceedings of ICASSP*, May 22–27, Prague, Czech Republic, pp. 5908–5911.

Saslove, H., and Yarmey, A. D. (**1980**). "Long-term auditory memory: Speaker identification," J. Appl. Psychol. **65**(1), 111–116.

Shue, Y.-L., Keating, P. A., Vicenik, C., and Yu, K. (**2011**). "VoiceSauce: A program for voice analysis," in *Proceedings of the ICPhS XVII*, Hong Kong, Vol. 126, pp. 1846–1849.

Simpson, A. P. (**2012**). "The first and second harmonics should not be used to measure breathiness in male and female voices," J. Phon. **40**(3), 477–490.

Smirnov, N. (**1948**). "Table for estimating the goodness of fit of empirical distributions," Ann. Math. Stat. **19**(2), 279–281.

Smith, H. M., Baguley, T. S., Robson, J., Dunn, A. K., and Stacey, P. C. (**2019**). "Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance," Appl. Cogn. Psychol. **33**(2), 272–287.

Stevenage, S. V., Tomlin, R., Neil, G. J., and Symons, A. E. (**2021**). "May I speak freely? The difficulty in vocal identity processing across free and scripted speech," J. Nonverbal Behav. **45**(1), 149–163.

Sun, X. (**2002**). "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 13–17, Orlando, FL, pp. I–333–I–336.

Van Leeuwen, D. A., and Brümmer, N. (**2007**). "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification I: Fundamentals, Features, and Methods* (Springer, Berlin-Heidelberg), pp. 330–353.

Wayland, R., and Jongman, A. (**2003**). "Acoustic correlates of breathy and clear vowels: The case of Khmer," J. Phon. **31**(2), 181–201.