Title

Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles.

Journal

The Journal of the Acoustical Society of America, 144(1)

ISSN

0001-4966

Authors

Park, Soo Jin
Yeung, Gary
Vesselinova, Neda
et al.

Publication Date

2018-07-01

DOI

10.1121/1.5045323

Peer reviewed

Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles

Soo Jin Park,[1,a)] Gary Yeung,[1] Neda Vesselinova,[2,b)] Jody Kreiman,[2,b)] Patricia A. Keating,[3] and Abeer Alwan[1]

[1]*Department of Electrical and Computer Engineering, University of California, Los Angeles, Los Angeles, California 90095, USA*

[2]*Department of Head and Neck Surgery, School of Medicine, University of California, Los Angeles, Los Angeles, California 90095, USA*

[3]*Department of Linguistics, University of California, Los Angeles, Los Angeles, California 90095, USA*

(Received 8 January 2018; revised 21 May 2018; accepted 17 June 2018; published online 26 July 2018)

Little is known about human and machine speaker discrimination ability when utterances are very short and the speaking style is variable. This study compares text-independent speaker discrimination ability of humans and machines based on utterances shorter than 2 s in two different speaking styles (read sentences and speech directed towards pets, characterized by exaggerated prosody). Recordings of 50 female speakers drawn from the UCLA Speaker Variability Database were used as stimuli. Performance of 65 human listeners was compared to i-vector-based automatic speaker verification systems using mel-frequency cepstral coefficients, voice quality features, which were inspired by a psychoacoustic model of voice perception, or their combination by score-level fusion. Humans always outperformed machines, except in the case of style-mismatched pairs from perceptually-marked speakers. Speaker representations by humans and machines were compared using multi-dimensional scaling (MDS). Canonical correlation analysis showed a weak correlation between machine and human MDS spaces. Multiple regression showed that means of voice quality features could represent the most important human MDS dimension well, but not the dimensions from machines. These results suggest that speaker representations by humans and machines are different, and machine performance might be improved by better understanding how different acoustic features relate to perceived speaker identity. https://doi.org/10.1121/1.5045323

[MIM] Pages: 375–386

I. INTRODUCTION

The human voice acts as a biometric that distinguishes individuals from one another, but it is a behavioral biometric, which makes it prone to variability. For example, various factors such as the speaker's mood, health condition, and speaking style influence the acoustic characteristics of speech sounds. This can lead to confusions for both humans and machines when distinguishing one individual from another (Hansen and Hasan, 2015; Kreiman and Sidtis, 2011; Schweinberger *et al.*, 2014). Thus, it is important to analyze human and machine performance in distinguishing speakers under such conditions of variability. This study investigates one of the most basic tasks in distinguishing speakers: deciding whether two speech samples came from a single speaker or from two different speakers. This task is referred to as *speaker discrimination* in human perception studies and as *speaker verification* in automatic speech processing studies. In particular, we focus on comparative effects of within-speaker variability in phonetic content and speaking style when utterances are very short (<2 s), and on the differences between perceptual and computational strategies that might account for performance differences.

A. Human speaker discrimination

For humans, discriminating unfamiliar voices is a separate decision-making process from recognizing familiar voices (Van Lancker and Kreiman, 1987). While familiar speaker recognition can be thought of as a gestalt-matching task, unfamiliar speaker discrimination additionally involves acoustic feature comparisons. This study uses unfamiliar speaker discrimination in order to analyze the acoustic features related to speaker identity. Several studies have shown that the perception of an unfamiliar voice requires both a generic speaker pattern that acts as a mental reference and a speaker-specific pattern that deviates from that reference (Kreiman and Sidtis, 2011, Chap. 5.3.4). Such a standard pattern, acquired over a lifetime, includes both how human voices generally sound and what aspects of speech are related to the speaker's identity.

Even though results vary widely depending on the experimental protocol used, humans are reasonably accurate at distinguishing unfamiliar speakers even with short utterances. For example, Kreiman and Papcun (1991) found that humans were 82.36% accurate in a speaker discrimination task with single-sentence (≈2 s) pairs. Human performance

[a)]Electronic mail: sj.park@ucla.edu

[b)]Also at: Department of Linguistics, University of California, Los Angeles, Los Angeles, CA 90095, USA.

generally improves as the utterance length increases until it plateaus with utterances longer than 60 s (Bricker and Pruzansky, 1966; Legge *et al.*, 1984). Authors disagree on why longer stimuli produce better results. Roebuck and Wilding (1993) found evidence supporting the hypothesis that the advantage of longer stimuli is in broader coverage of phonetic content. However, Cook and Wilding (1997) argued that the critical factor was not the number of different sounds heard, but rather the duration of the utterances due to speaker-specific prosody, speaking rate, and other non-phonemic aspects of the speech signal that are present in longer utterances (Kreiman and Sidtis, 2011, Chap. 7.3.1).

The effect of speaking style variability on human speaker discrimination has not been studied extensively. Studies in forensic speaker identification note that a speaking style mismatch between a criminal's voice heard at a crime scene and speech samples collected in a voice lineup (e.g., shouting versus reading) might confuse earwitnesses (see Jessen, 2008). In a speaker discrimination context, we expect such speaking style variability to cause a significant performance degradation based on results from a few studies dealing with emotion variability (e.g., Saslove and Yarmey, 1980). In that study, when a target voice changed tone (related to emotion or affect), mean "hit-miss" and "false alarm-correct rejection" scores decreased significantly.

B. Machine speaker verification and how it compares to human speaker discrimination

State-of-the-art automatic speaker verification (ASV) systems are typically pre-trained with large amounts of data from a large number of speakers. Hundreds of hours of recordings are used to train a statistical model for human speech, called a *universal background model* (UBM, Reynolds *et al.*, 2000). A widely-used feature set for the statistical modeling is based on mel-frequency cepstral coefficients (MFCCs), which approximate the spectral envelope of the speech signal. A new utterance can be thought of as a deviation from the UBM. The nature and extent of the deviation, however, will be influenced by both speaker-specific and utterance-specific information. Thus, these systems need to minimize within-speaker variability while maximizing between-speaker variability. Hundreds of additional hours of recordings are used to train a subspace onto which the deviation is projected. The projected low-dimensional vector, referred to as an *i-vector* (Dehak *et al.*, 2011), is thought to represent speaker identity. When the system receives a pair of speech samples as inputs, an i-vector is found for each utterance. Then, the likelihood that the i-vectors represent the same speaker is calculated based on the pre-trained model and subspaces. Probabilistic linear discriminant analysis (PLDA, Kenny *et al.*, 2013) is often used to calculate this likelihood. The system then applies a threshold to the likelihood to make a same versus different speaker decision.

Automatic speaker discrimination can be viewed as analogous to perceptual speaker discrimination, although the latter is much more complicated than statistical pattern recognition based on frame-level features. That is, the pre-trained UBM and subspaces are analogous to a human's pre-existing idea of the average speaker model and the manner in which a new voice differs from it. Such a model represents the life-long experience of the listener with voices, with internal structuring that is not yet understood. Despite this analogy, however, differences presumably exist between the speaker-distinguishing strategies used by humans and machines as evidenced by poor machine performance for difficult tasks. Challenging conditions include very short utterances (<2 s), text-independent tasks, and speech spoken in different styles.

Although machines outperform humans on long utterances in certain conditions (e.g., Hautamäki *et al.*, 2010; Kahn *et al.*, 2011), their performance on short utterances is seemingly worse than that of humans. For example, a state-of-the-art text-independent ASV system using MFCCs was 97.60% accurate at discriminating speakers with 2.5-min-long pairs, but it was only 89.48% accurate with 5-s-long pairs, and performance worsened to 77.69% accuracy with 2-s-long pairs on the National Institute of Standards and Technology (NIST) speaker recognition evaluation (SRE) 2003 database (Das and Prasanna, 2016), compared to 82.36% for humans hearing short utterances, as noted earlier. Similar performance degradation was observed in our previous study on the SRE 2010 database, which showed 97.11%, 83.10%, and 71.53% accuracy for 2.5-min, 5-s, and 2-s-long pairs, respectively (Park *et al.*, 2017). As mentioned above concerning human performance, one reason for the degradation with shorter utterances could be that there is insufficient phonetic coverage for the machines to infer appropriate statistics. Text dependency also affects machine performance. For example, when utterances are short (<10 s), matching phonetic content by using same-text pairs yields error rates that are approximately half those of the text-independent pairs (Das *et al.*, 2016; Park *et al.*, 2017). One exception occurs when short digit sequences (<2 s) are used. In that limited-vocabulary case, performance can reach 95% accuracy or higher (Larcher *et al.*, 2014).

Although the effect of speaking style mismatch has not yet been studied extensively in ASV communities, some studies on emotion variability are available. For example, Parthasarathy *et al.* (2017) reported that an emotion mismatch between utterances degraded ASV system performance, which worsened as the utterance length decreased from 11 to 2.75 s for naturalistic (not acted) expressive voices. However, because that study did not compare matched emotion conditions, the amount of degradation that can be attributed to emotion variability is not clear. In Nakasone and Beck (2001), it was noted that the performance of a speaker identification system degrades when trained with spontaneous speech and tested on reading, compared to when spontaneous speech was used for both training and testing, even though the utterances were long (29 s). The system was a closed-set speaker identification task, which is not directly comparable to speaker verification tasks, but it is expected that ASV performance might also decrease due to speaking style differences.

Considering that performance degrades in text-independent ASV mainly due to the sensitivity of MFCCs to phonetic content, various features that are thought to be less sensitive to such variability might improve system performance

(Das and Prasanna, 2017). For example, Das and Prasanna (2016) used features derived from the linear prediction residual signal to represent voice source characteristics. These features improved the system performance by providing additional or complementary information to conventional cepstral features on text-independent tasks when the speakers were modeled with 2.5-min-long utterances and tested with short utterances (2–10 s). Other studies have shown that the phase components of a speech signal are important for speaker identity (Vijayan *et al.*, 2016), and such information could be used for text-independent long-utterance ASV. Approaches to capturing speaker-specific prosody have also been proposed (Dehak *et al.*, 2007; Reynolds *et al.*, 2003; Shriberg *et al.*, 2005). The effectiveness of phase and prosodic features evaluated on long-utterance (>1 min) text-independent ASV in these studies suggests that they might provide additional speaker information for short-utterance tasks as well.

In our recent work (Park *et al.*, 2016; Park *et al.*, 2017), we used voice quality features that were inspired by a psychoacoustic model of voice perception that accounts for perceived voice quality (Kreiman *et al.*, 2014). This set of features was applied to short-utterance (<2 s) text-independent ASV, and it successfully improved system performance by providing complementary information to conventional cepstral features.

If humans are more accurate than machines in distinguishing speakers from short utterances that include large phonetic variability, understanding speaker perception strategies might help improve machine performance. The above-mentioned studies provide general ideas about how well humans and machines perform under various conditions. However, to our knowledge, a direct and detailed comparison between human and machine speaker discrimination under conditions of within-speaker text- and speaking-style variability and using very short utterances has not yet been made, in part because we lack proper databases to undertake such studies. Recently, a database was developed at the University of California, Los Angeles (UCLA) to represent both within- and between-speaker variability and recording session variability (see Keating *et al.*, 2018; Kreiman *et al.*, 2015). The UCLA Speaker Variability Database includes a large number of speakers (currently 103 female and 105 male students at UCLA, aged 18–37; mean age 20), with multiple recording sessions and varying phonetic content, speaking style, and affect conditions per speaker, reflecting normal, daily-life variations in voice quality. For the present study, very short speech samples (<2 s) with high cross-task variability were selected to probe the limits of both humans and machines when confronted with high within-speaker variability.

This study addresses the following questions: (1) how much do human and machine performance degrade when utterances are very short (<2 s), phonetic content varies and style variability is large? (2) What is the performance gap between humans and machines in such conditions? (3) What is the difference in the features and strategies used by humans and by machines? The main focus is on exploring the effect of including psychoacoustically-valid acoustic indices of voice quality in ASV, given that these features specify voice quality for human listeners.

The rest of the paper is organized as follows. In Sec. II, we describe the databases used in this study. In Sec. III, we present human speaker discrimination results, while in Sec. IV we describe analyses of the performance of ASV systems on the same stimuli given to human listeners. In Sec. V, human and machine performances are compared, and the differences between human speaker discrimination and machine speaker verification are analyzed. The paper concludes with a summary and conclusion in Sec. VI.

II. DATABASES

A. Voice samples from the UCLA speaker variability database

Voice samples were drawn from the UCLA Speaker Variability Database (see Keating *et al.*, 2018; Kreiman *et al.*, 2015, for more information). Speakers were recorded in a sound-attenuated booth on three different days. All speech was elicited via on-screen displays and recorded using a 1/2 in. Brüel & Kjær microphone with a sampling rate of 22 kHz and a fixed mouth-to-microphone distance. The speech samples used in this study were later resampled to an 8 kHz sampling rate so that speech bandwidth would equal that of the SRE databases used for training the ASV systems.

The present study used speech from two of the tasks included in the database: read sentences and pet-directed speech. These two speaking styles are the most distinct in the database. Read sentences are text-constrained clear speech, while pet-directed speech is spontaneous and includes exaggerated prosody. Speakers produced two repetitions of five Harvard sentences (IEEE Subcommittee on Subjective Measurements, 1969) in all three recording sessions, for a total of six repetitions of each sentence and 30 sentences overall. The sentences were "The boy was there when the sun rose," "Kick the ball straight and follow through," "Help the woman get back to her feet," "A pot of tea helps to pass the evening," and "The soft cushion broke the man's fall." For pet-directed speech, speakers were instructed to talk to pets displayed in a video. They could choose between a kitten video (2 min 36 s) and a puppy video (1 min 51 s). Resulting utterances were often (but not always) characterized by exaggerated prosody, similar to infant-directed speech (Burnham *et al.*, 2002).

Fifty female self-reported native speakers of English were randomly selected from the database for use in the experiments reported in this study. Female speakers were chosen because they used more prosodic exaggeration when talking to pets than did male speakers, leading to larger differences between the read sentences and the pet-directed speech. *Post hoc* listening by two of the authors indicated that utterances from nine speakers were perceptually "marked" by a non-American dialect, overly-precise articulation and/or unusual disfluencies in reading. The remaining 41 speakers lacked such personal idiosyncrasies and will be referred as "unmarked."

B. NIST SRE database

While the UCLA Speaker Variability Database provided all the evaluation utterances for the present study, separate

speech databases were used to pre-train the ASV systems tested here. The SRE databases developed by NIST are often used to train a UBM and speaker variability subspaces. We used the NIST SRE04, 05, 06, and 08 databases (Martin and Greenberg, 2009; Przybocki and Martin, 2004; Przybocki *et al.*, 2006) for this purpose. These databases provide more than 3000 h of speech samples from 2692 female and 1115 male speakers, and have a variety of channels including telephone speech, microphone, and "interview" speech.

Because the evaluation utterances were all from female speakers, only the recordings from female speakers were used to train the UBM and subspaces. In addition, evaluation recordings were downsampled to an 8 kHz sampling rate to match the bandwidth of the SRE databases.

III. HUMAN PERCEPTION EXPERIMENTS

We first tested human listeners' ability to discriminate among the speakers across the two speaking styles (read sentences and pet-directed speech). Recall that we employed this high cross-task variability to probe the limits of performance for humans and machines.

A. Method

For each speaker, three read sentences were selected from each of the three recording sessions. Each speech sample lasted less than 2 s. Two excerpts were taken from the pet-directed speech, matched in length to the average duration of the sentences. These stimuli, downsampled as described above, were assembled into 100 pairs of voices in which both voice samples came from the same person (50 pairs of read sentences and 50 pairs where a read sentence was paired with pet-directed speech), and 2450 pairs where the two speakers were different (half including two read sentences and half including one read sentence and one pet-directed speech sample), for a total of 2550 pairs of stimuli. Stimuli were always drawn from different recording sessions, and each pair included two different read sentences. Thus, this task was always text- and recording session-independent.

To minimize listener fatigue, stimuli were divided at random into 12 subsets of 200 pairs of voices and 1 subset of 150 pairs. Thirteen groups of five normal-hearing UCLA students and staff members (aged 18–28; mean age 19.91; standard deviation 2.28; 65 listeners total) were recruited, of whom 30 considered themselves L1 English speakers. The participants listened to the pairs of stimuli over Etymotic insert earphones (model ER-1) at a comfortable constant listening level. Each pair could be played only once in each presentation order (AB/BA). The listeners were asked whether the two speech samples were produced by the same speaker or by two different speakers. They also reported their confidence in their responses on a 1–5 scale (1 = positive, 5 = wild guess). They were not told how many speakers were represented in the trials. The experiment was self-paced, and listeners were encouraged to take breaks as needed. Total testing time was less than one hour per listener.

B. Results

Hit rates (HRs) and false alarm (FA) rates were calculated by defining a hit as a correct "same speaker" response and a false alarm as an incorrect "same speaker" response. Additionally, listeners' same versus different responses were combined with their confidence ratings to create a scale ranging from "positive, same speaker" (= 1) to "positive, different speakers" (= 10). These scalar responses were used to derive receiver operating characteristic (ROC) curves using SYSTAT software (Systat Software Inc., 2018). d' (d-prime, e.g., Macmillan and Creelman, 2005) and the area under the receiver operating characteristic curve (AUC) were calculated for each ROC curve. Note that d' values calculated from ROC curves can differ from values directly calculated from hit and false alarm rates. The equal error rate (EER) was computed from the ROC curve derived from listeners' confidence ratings, because humans do not have full control of their decision threshold and their EERs cannot be calculated directly.

Hit rates, false alarm rates, d' (from the ROC curve), AUC, and EER are shown in Table I. Because listener performance could be affected by the speakers' perceptual markedness, results when the stimuli were pairs from the 41 unmarked speakers, pairs from the nine marked speakers, pairs consisting of one marked and one unmarked speaker, and pairs from all 50 speakers are shown separately in the

TABLE I. Composite human speaker discrimination performance for the 41 perceptually-unmarked speakers, nine perceptually-marked speakers, pairs consisting of one marked and one unmarked speaker, and all 50 speakers in terms of HR (%), FA (%), d', the AUC, and the EER (%). Read-read and read-pet indicate that the token pair presented to the listener was composed of two different read sentences or one read sentence and one pet-directed speech segment, respectively. Note that there were no "same speaker" pairs when listeners compared one marked speaker to one unmarked speaker, so that the hit rate could not be calculated.

	No. same speaker pairs	No. different speaker pairs	HR	FA	d'	AUC	EER
Read-read, unmarked speaker pairs	41	820	87.3	25.8	1.81	0.885	19.02
Read-pet, unmarked speaker pairs	41	820	54.1	35.2	0.50	0.644	39.23
Read-read, marked speaker pairs	9	36	68.9	21.7	1.48	0.844	24.86
Read-pet, marked speaker pairs	9	36	37.8	34.4	0.16	0.538	46.23
Read-read, marked/unmarked pairs	N/A	369	N/A	20.7	N/A	N/A	N/A
Read-pet, marked/unmarked pairs	N/A	369	N/A	32.1	N/A	N/A	N/A
Read-read, all speaker pairs	50	1225	84.0	24.2	1.73	0.876	20.19
Read-pet, all speaker pairs	50	1225	51.2	34.2	0.46	0.628	40.34
All pairs	100	2450	67.6	29.2	1.11	0.766	30.58

table. The pairs of read sentences are denoted as "read-read" and the pairs of one read sentence and one pet-directed speech excerpt are denoted as "read-pet."

Human listeners were reasonably accurate in distinguishing unmarked speakers when stimuli were pairs of read sentences ($d' = 1.81$). As expected, accuracy decreased when listeners heard read speech paired with pet-directed speech ($d' = 0.50$). Changes in hit and false alarm rates were similar in magnitude (38.03% and 36.43%, respectively), suggesting that results reflect a difference in discriminability without an accompanying change in response biases. Because there were many more unmarked speakers than marked speakers, the "all speaker pairs" results are very similar to those for the unmarked speakers for the read-read pairs.

Although the marked speakers had idiosyncrasies in their speech, they were in fact harder to discriminate. d' equaled 1.48 for read-read pairs (compared to 1.81 for the unmarked speakers), and 0.16 for read-pet pairs (compared to 0.50). The performance degradation reflected a large decrease in the hit rate and a smaller decrease in the false alarm rate, suggesting a stricter response criterion. For trials including one marked speaker and one unmarked speaker, only false alarm rates could be calculated because stimuli always came from different speakers. Those marked/unmarked pairs had the lowest observed false alarm rates: 20.7% for read-read pairs and 32.1% for read-pet pairs.

C. Discussion

Humans were reasonably accurate in distinguishing speakers from read-read pairs, consistent with results from other studies (e.g., Kreiman and Papcun, 1991). In contrast, human speaker discrimination accuracy decreased considerably for read-pet pairs, with d' less than 1.0 for all such comparisons. One issue for these pairs might have been the limited phonetic content of the pet-directed speech excerpts. While the read sentences were phonetically rich, pet-directed speech was largely limited to phrases such as "Awww, cute," with stereotyped intonation contours that lacked the idiosyncrasies of the read-read pairs.

Moreover, there is a significant difference in $F0$ between the read sentences and pet-directed speech. The mean $F0$ for the read sentences was 221.23 Hz, while that of pet-directed speech was 313.02 Hz [$F(1, 548) = 575.2$, $p < 0.01$]. The extraordinarily high $F0$ of the pet-directed speech might have confused listeners, who typically rely heavily on $F0$ when assessing speaker identity (Baumann and Belin, 2010; Nolan *et al.*, 2011). Additionally, exaggerated prosody makes other cues, such as pauses between words and speaking rate, sound different from read sentences.

Differences in perception when listening to marked versus unmarked speakers emphasize the importance listeners place on specific cues, such as an unfamiliar accent or disfluency, even when stimuli are short (<2 s). Note that speakers' word choice was not a cue in this experiment, because the sentences were given and the pet-directed speech did not include much lexical variety. In this context, decreases in performance when speakers were perceptually marked is consistent with previous findings that accented speakers are more difficult to identify than unaccented speakers, especially when the utterances are short (<1 s) (Goldstein *et al.*, 1981), and that listeners are better at discriminating among speakers when they are familiar with the phonetic inventory used by particular speakers (Kreiman and Sidtis, 2011, Chap. 7.2.3). Responses to the speech of the marked speakers were not only less accurate, but may also have been biased to "different speaker" decisions, possibly because listeners had difficulty distinguishing features specific to the speaker from features that characterized differences in phonetic content or speaking style between utterances.

IV. ASV EXPERIMENTS

This section describes application of an i-vector/PLDA ASV system to the stimuli just described. The same tasks presented to the human listeners were given to the ASV system, permitting a fair comparison between humans and machines.

A. Feature extraction

Performance of ASV systems depends, in part, on the use of appropriate features to distinguish speakers. The feature sets used in the ASV experiments are discussed in this subsection. All features were automatically extracted, and no manual refinements were made.

1. MFCCs

MFCCs of dimension 20 were calculated every 10 ms using a 25-ms-long window. The coefficients and their first derivatives were used as a feature set. Second derivatives were not used because they did not provide notable performance gain in our preliminary work.

2. VQual2: Voice quality features

In this section, we describe a novel set of features inspired by a psychoacoustic model of voice quality (Garellek *et al.*, 2016). This feature set comprised $F0$, $F1$, $F2$, $F3$, cepstral peak prominence (CPP, Hillenbrand *et al.*, 1994), and three measures of source spectral slope. The slope features were generated by estimating the amplitudes of the first, second, and fourth harmonics and of the harmonic nearest to 2 kHz (denoted $H1$, $H2$, $H4$, and $H2k$), and then calculating the differences between them. Amplitude difference features were denoted as $H1^*$-$H2^*$, $H2^*$-$H4^*$ and $H4^*$-$H2k^*$, where the asterisks (*) indicate that harmonic amplitudes were corrected for the effects of formant frequencies on amplitude (Hanson, 1997; Iseli *et al.*, 2007).[1] The features were extracted pitch-synchronously every 10 ms. The effectiveness of this initial feature set, referred to as VQual1, on ASV was tested in our previous study (Park *et al.*, 2016). The feature set was later modified to better represent speaker identity for ASV (Park *et al.*, 2017). The modification was based on the f-ratio criterion (Lu and Dang, 2008; Nicholson *et al.*, 1997), which measures how well an individual feature separates classes of stimuli. This criterion is widely used to identify features which have large between-class variance and small within-class variance:

$$f = \frac{\text{between class variance}}{\text{within class variance}} = \frac{\frac{1}{M}\sum_{i=1}^{M}(\mu_i - \mu)^2}{\frac{1}{M}\sum_{i=1}^{M}\sigma_i^2}, \quad (1)$$

where M is the number of classes, μ_i is the within-class mean, μ is the global mean, and σ_i^2 is the within-class variance of a single feature.

In Park *et al.* (2017), read sentences and pet-directed speech samples from 100 female and 100 male speakers in the UCLA database were analyzed using the *f*-ratio with a large number of features. Although the *f*-ratio results were different between the two speaking styles, feature ranks were similar. Thus, a modified feature set denoted as VQual2 was constructed, including *F*0, *F*1, *F*2, *F*3, *H*1–*H*2, *H*2–*H*4, *H*4–*H*2k (without formant correction[2]), formant amplitudes *A*1, *A*2, *A*3, and CPP. Note that the original VQual1 feature set was generated from a psychoacoustic model of voice quality, but the modified VQual2 set was chosen to maximize ASV performance. The variation in the feature sets might be partly due to the difficulties in automatic measurement and/or large within-speaker variance. It might also be due to the fact that VQual1 was evaluated on sustained vowel sounds while the new feature set was evaluated on continuous speech signals.

B. Method

An i-vector/PLDA ASV system was used to analyze machine performance. The i-vector size was 600 and it was reduced to 200 after the PLDA. The UBM (modeled with 2048 Gaussians) and subspaces were trained with female voices using the NIST SRE databases. The two feature sets described above, MFCCs and VQual2, were used in the experiments.[3]

After obtaining the PLDA scores from each system, score fusion was performed to test for further improvements (Ramachandran *et al.*, 2002). Fusion is analogous to averaging human listeners' dissimilarity scores and making a new decision based on the average score. The fusion system outputs were linearly combined using the following equation:

$$s = \alpha s_v + (1 - \alpha)s_m, \quad (2)$$

where s_m is the PLDA score using MFCCs, s_v is the PLDA score using VQual2 features, and α, the coefficient of s_v, ranges from 0 to 1. PLDA scores using both MFCCs and VQual2 features were scaled to have zero-mean and unit-variance before the linear combination was performed. The coefficient α was set to 0.452 so that it yields the lowest EER for the condition composed of all possible pairs.

C. Results and discussion

The AUC and the EER were calculated to measure system performance. AUC values were estimated using SYSTAT software to facilitate comparisons with human performance. Machine and human results are shown in Table II. In general, score fusion improved machine performance. For read-read pairs using all speakers, for example, the AUC for the MFCC feature set, VQual2 feature set, and for the fusion of the two were 0.776, 0.683, and 0.791, respectively. Thus, while performance of VQual2 alone does not exceed the performance of MFCCs, fusing the two systems seemingly provided complementary information that improved performance. Other studies have also shown that fusing complementary features improves ASV performance for 10-s utterances (Das and Prasanna, 2017). This pattern was observed in most of the other comparisons. The exceptions where the fusion resulted in a slight performance degradation were for read-read pairs from marked speaker pairs (from 0.687 to 0.683), and when all pairs were combined (from 0.716 to 0.714).

The decrease in performance of the VQual2 features due to style mismatches was smaller than that observed for MFCCs, although overall performance was generally worse for VQual2 features. For unmarked speakers, the EER for VQual2 increased from 36.08% for read-read pairs to 44.09% for read-pet pairs (22.20% relative decrease in performance), where the EER for MFCCs increased from 30.31% to 44.17% (45.73% relative decrease in performance). For marked speakers, the VQual2 EER increased from 41.58% to 44.91% (8.01% relative decrease in performance), while the MFCC EER increased from 32.03% to 39.31% (22.73% relative decrease in performance).

The robustness to style variability suggests that voice quality features might be effective for conditions that are challenging to conventional cepstral features. Note, however, that our previous study (Park *et al.*, 2017) found that

TABLE II. ASV performance evaluated using the same stimuli as in the human perception experiments. The AUC was measured, and the EER (%) was calculated from the ROC curve. Human perception results in terms of AUC and EER are repeated from Table I in the last column for comparison. The best performance for each condition is boldfaced.

	MFCC		VQual2		fusion		human	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Read-read, unmarked speaker pairs	0.765	30.31	0.679	36.08	0.780	29.21	**0.885**	**19.02**
Read-pet, unmarked speaker pairs	0.587	44.17	0.581	44.09	0.601	47.54	**0.644**	**39.23**
Read-read, marked speaker pairs	0.687	32.03	0.657	41.58	0.683	31.78	**0.844**	**24.86**
Read-pet, marked speaker pairs	0.593	39.31	0.531	44.91	**0.601**	**37.35**	0.538	46.23
Read-read, all speaker pairs	0.776	29.17	0.683	36.18	0.791	28.71	**0.876**	**20.19**
Read-pet, all speaker pairs	0.594	43.44	0.587	43.55	0.615	42.79	**0.628**	**40.34**
All pairs	0.716	35.97	0.627	43.18	0.714	36.52	**0.766**	**30.58**

the performance degradation of the VQual2 features due to style mismatches was similar to or worse than that of MFCCs. Unfortunately, a direct comparison with that study is not appropriate because the speech samples used in that study were 5-s long while the speech samples in this study were less than 2-s long. Since longer utterances benefit both MFCC and VQual2 feature sets, especially if the phonetic content is richer, it might be the case that the advantage of having more phonetic content in read sentences outweighed the within-speaker variability in speaking style. Thus, the compounded effect of utterance length, phonetic content and style variability requires further analysis for machines. In addition, Park *et al*. (2017) did not consider speaker markedness, which might also have been a factor impacting system performance.

Unexpectedly, MFCC performance for read-pet, marked speaker pairs (EER = 39.31%) was better than that for unmarked speakers (EER = 44.17%). For VQual2, the performance degraded, but the difference was small (from 44.09% to 44.91%). The effect of markedness on machine performance will be analyzed with a larger number of speakers in a follow-up study.

Note that the AUC and EER measures are not always correlated because the two measures reflect different properties of the curve. The AUC is calculated from the entire ROC curve, and it reflects overall accuracy regardless of a specific decision threshold. On the other hand, the EER only focuses on the point where the false rejection rate and the false acceptance rate are equal, and it summarizes the system performance in terms of the error rate. These measures can differ, especially when the ROC curves are skewed. Skewed ROC curves can result when the variance of the distribution in the decision space of same-speaker pairs is different from that of the different-speaker pairs (Macmillan and Creelman, 2005). While the EER is a widely used metric for machines, for humans this metric might be misleading because humans cannot consciously adjust their decision threshold. Thus, in the rest of the paper, the AUC is used to compare human and machine performance, and the EER is used only to compare machine performance in different conditions.

V. COMPARING HUMAN AND MACHINE SPEAKER DISCRIMINATION

This section compares the human and machine speaker discrimination results in the face of within-speaker variability as presented in Secs. III B and IV C. The purpose of the comparison is to investigate performance differences between humans and machines when large within-speaker variability makes the task difficult for both, and to analyze the factors that affect performance. Recall that all tasks are text- and recording-session-independent.

A. Overall performance comparison

Human and machine performances are compared in Table II. Humans performed better than machines in most conditions. For instance, with unmarked speakers, the AUC for ASV score fusion was 0.780 for read-read pairs, compared to AUC = 0.885 for humans.

Performance differences between humans and machines could be due to many factors. First, humans can utilize multiple levels of information from the audio signal, but machines rely on frame-level features. For example, humans routinely attend to individual speakers' unique prosody, idiosyncrasies in voice onset time, and so on, but ASV systems consider the distribution of features extracted from 25-ms frames and at most their time derivatives. Second, it is likely that even when humans and machines use similar acoustic information, they process the information in different ways to make their same versus different speaker decisions.

For read-read pairs, machines were less robust to markedness than humans were. Fusion performance on read-read pairs from unmarked speakers resulted in an AUC of 0.780, while the AUC for marked speakers equaled 0.683 (12.44% relative decrease in performance). Human performance resulted in AUCs of 0.885 and 0.844 (1.24% relative decrease in performance) for the unmarked and marked speakers, respectively. Because the UBM represents the overall smoothed distribution of the acoustic features from a large number of speakers, idiosyncrasies due to speaker markedness might not be well-represented with this model. In addition, if similar idiosyncrasies are not well-represented in the pre-training data, the machine will fail to model the between-speaker variability from these idiosyncratic differences, leading in turn to performance degradation.

On the other hand, machines were more robust to markedness for read-pet pairs than were humans. Fusion AUCs for read-pet pairs from unmarked speakers and from marked speakers were both 0.601. However, the AUC for human listeners decreased from 0.644 (unmarked speakers) to 0.538 (marked speakers), a 16.46% relative decrease in performance. Even though it might be difficult to generalize because there were only nine marked speakers, these results imply that machines are less sensitive to speaker markedness than humans are when the acoustic characteristics of the speech change due to prosody exaggeration. The compound effect of speaker markedness and speaking style on human and machine performance can be explored in the future by including recordings from L2 English speakers.

It was consistently observed that the performance gap between humans and machines was smaller for mismatched speaking styles. For instance, with read-pet, unmarked pairs, the AUC for fusion was 0.601 and the AUC for humans was 0.644, while the AUCs for the read-read, unmarked pairs was 0.780 for fusion and 0.885 for humans. The interesting small performance gap between humans and machines for the read-pet condition will be analyzed in detail in future studies.

B. Performance analysis for subsets of a smaller number of speakers

Previous studies (Kreiman and Gerratt, 1996) have shown that listener performance in discrimination tasks is characterized by flexible, idiosyncratic perceptual strategies, such that a feature may be important for distinguishing some pairs of speakers but not others. Given this situation, combining too many speakers in a single analysis obscures the strategies used by listeners because relations in the "perceptual

speaker space" become too complicated to summarize even with a large number of parameters. For this reason, we conducted further analyses using small ($n = 15$) subsets of the original set of 41 unmarked speakers. The analyses were restricted to read-read sentence pairs, because the main purpose is to investigate the difference in decision strategies between humans and machines, and performance between humans and machines differed most for these pairs. With 15 speakers, each subset had 15 same-speaker pairs and 105 different-speaker pairs. Ten sets of 15 speakers were randomly selected from the 41 unmarked speakers. Three of the ten subsets (RAND1, RAND2, and RAND3) were chosen for multi-dimensional scaling (MDS) analysis so that each unmarked speaker was included in at least one of the subsets. Discrimination data for the read sentence pairs used in the perception experiment were extracted, and the performances of humans and machines were calculated for each subset.

As shown in Table III, the AUC for human listeners varied between 0.851 and 0.909, and the EER varied between 16.10% and 21.53%. MFCC performance was worse than humans' and more variable across subsets: its AUC varied between 0.679 and 0.772, and the EER varied between 26.61% and 40.57%. The AUC for score fusion varied between 0.713 and 0.792, and the EER varied between 24.18% and 34.02%. VQual2 performance was most consistent (although not best) among the three ASV systems; its AUC ranged from 0.678 to 0.684, and its EER ranged from 34.44% to 36.75%.

The three subsets showed different rankings of the three ASV systems. In RAND1, the MFCC system had a much better EER (26.61%) than VQual2 (36.75%), and fusion showed the best performance (24.18%). In RAND2, MFCC performance (30.50%) was better than that of VQual2 (34.44%) and was similar to fusion (30.31%). In RAND3, VQual2 performance (36.35%) exceeded MFCC (40.57%) and was improved by fusion (34.02%).

C. MDS analysis

Nonmetric MDS (Kruskal and Wish, 1978) was applied to provide insight into the differences in the information utilized by humans and machines underlying these results.

TABLE III. Human and machine performance in terms of EER (%) and the AUC. The performance is measured within the 10 subsets of 15 randomly selected speakers reading sentences. The mean and standard deviation across the ten subsets are shown in the first two rows. Performance on three of the ten subsets (RAND1, RAND2, and RAND3) used for MDS analysis is shown in the rest of rows. There were 15 same-speaker pairs and 105 different-speaker pairs in each subset. Fusion indicates that a linear score fusion is used between the MFCC and VQual2 systems.

	MFCC		VQual2		fusion		human	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Mean	0.726	32.99	0.680	35.87	0.744	29.82	0.890	18.49
Standard deviation	0.056	6.21	0.040	2.65	0.039	3.80	0.021	2.42
RAND1	0.772	26.61	0.680	36.75	0.792	24.18	0.851	21.53
RAND2	0.703	30.50	0.678	34.44	0.722	30.31	0.898	17.61
RAND3	0.679	40.57	0.684	36.35	0.713	34.02	0.909	16.10

MDS is often used in forensic studies to objectively measure perceived speaker similarity to construct fair voice lineups (McDougall, 2013). The MDS space can be thought of as a "(perceptual) speaker space" where the stimuli are close if they are perceived as similar. The MDS axes can be interpreted by examining correlations between the coordinates of the stimuli and acoustic or other measures of those stimuli: a high correlation suggests the measure might be an important cue for distinguishing speakers.

1. MDS space determination

For each 15-speaker subset, confidence ratings from human listening data were combined with same versus different judgements, such that a value of 1 (positive, same speaker) was assumed to mean the voices were very similar, and a value of 10 (positive, different speakers) meant they were maximally dissimilar. These scores were averaged across listeners and assembled into lower-half dissimilarity matrices. For the three ASV systems (MFCC, VQual2, and fusion), the dissimilarity between a pair was calculated as the negated PLDA score. Nonmetric MDS was then performed on the human data and on the ASV systems for each speaker subset. MDS solutions were calculated in 2–5 dimensions for each subset of the data, and solutions were chosen by reference to plots of the number of dimensions extracted versus R^2 and stress (Kruskal and Wish, 1978, pp. 48–60). R^2 measures the variance in dissimilarities explained by the MDS solution, and stress measures the overall fit of the scaling model to the data. Solutions were chosen based on elbows in plots of stress and R^2 versus the number of dimensions (Fig. 1). A four-dimensional solution best fits the human data for RAND1, while the solutions are three-dimensional for RAND2 and RAND3.

2. Relationship between human and machine decision spaces

Canonical correlation analysis (CCA, e.g., Tabachnick and Fidell, 2013) was used to evaluate the extent to which human and machine speaker spaces were related. Here, one set of the variables is the MDS coordinates from each of the three ASV systems (MFCC, VQual2, and fusion) for a speaker subset, and the other is the MDS coordinates from human responses for the same subset.

The resulting R^2 scores using three-component CCA are shown in Table IV. Dimensions of the machine MDS spaces were insufficiently interpretable in terms of the dimensions of the human perceptual space, suggesting that machines and humans used different strategies to discriminate speakers. For RAND1, at most 56.3% of the variance in the ASV speaker space was explainable using the dimensions from the speaker space derived from perceptual data. For RAND2/VQual2, the negative R^2 value indicates that the estimated model was worse than the constant model. The overall low R^2 values suggested that there was little relationship between human and machine speaker spaces, at least when a linear model was used.

If we compare the CCA results in Table IV with the EER performance in Table III, we notice that the relationship between the model fit and system performance was

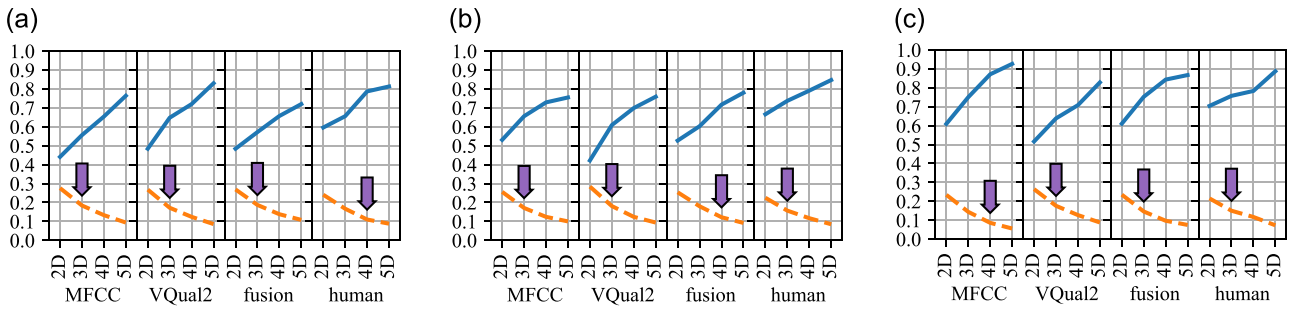

FIG. 1. (Color online) Calculated R^2 values (solid line) and stress (dashed line) for the MDS solutions in the (a) RAND1, (b) RAND2, and (c) RAND3 subsets for human data and for the ASV systems using MFCC, VQual2 features, and their score fusion. Arrows point to the elbow in each curve.

weak. For example, the best performing ASV system in Table III was fusion for all subsets. Fusion showed the highest R^2 value for RAND1 and RAND2, but not for RAND3. In addition, even though the R^2 value of RAND3 MFCC was the second highest ($R^2 = 0.503$), its performance was the worst (EER = 40.57%) among all three subsets.

3. Acoustic correlates of MDS axes

The weak relationship between human and machine speaker spaces suggests that acoustic information is used rather differently in human versus machine decision making. To examine this hypothesis, we analyzed how the 11 VQual2 acoustic measures were correlated with the MDS spaces for both sets of similarity data. The mean of each of the 11 acoustic measures was calculated for each utterance from all 50 speakers, and a factor analysis of dimension 5 was undertaken to reduce the number of predictor variables. A similar procedure was applied to the standard deviations of the acoustic measures. The absolute factor loadings, which reflect the correlations between the acoustic measures and factors, are shown in Fig. 2. For the acoustic means, factor 1 was mostly related to the formant amplitudes, and factor 4 showed a strong relationship with $F0$. For the standard deviations, factor 1 was highly correlated with formant amplitudes, and factor 2 was related to fundamental frequency, and to the first and the second formant frequencies.

Next, factor scores calculated at the utterance level were averaged within each speaker, after which we constructed five-dimensional acoustic speaker spaces for each subset. Finally, the relationship between the acoustic space and the MDS spaces was analyzed using multiple regression. Results are shown in Table V. Interestingly, factors estimated from the means of the acoustic measures were related to the most important dimension ($D1$) of the perceptual speaker spaces for all subsets, and the factors from the standard deviations, which can be related to the within-utterance variability, were related $D1$ of the MFCC speaker spaces for all subsets. For humans, factors 4 and 5, derived from mean acoustic measures, were statistically significant ($p < 0.05$) for the multiple regression model in RAND1 and RAND2, and RAND1 and RAND3, respectively. Recalling that factor 4 was highly related to $F0$ and factor 5 was related to $F3$, these results are consistent with previous studies that reported $F0$ and $F3$ being the most important acoustic predictors of human judgements (e.g., Baumann and Belin, 2010; Nolan *et al.*, 2011). In RAND2, factor 2 from the mean data, which was related to $F2$, $F3$, and $A3$, was significantly related to human $D1$, and in RAND3, factor 1 from the standard deviation data, which was related to formant amplitudes, was significantly related to human $D1$. These results suggest that formant amplitudes might also provide important information for human decision-making.

For MFCCs, factor 5 from the standard deviation data was significantly related to $D1$ for subsets RAND1 and RAND3. For VQual2, which was derived from a psychoacoustic model

TABLE IV. R^2 scores of the canonical correlation analysis between the MDS space from the three ASV systems (MFCC, VQual2, and fusion) and human MDS space in each speaker subsets (RAND1, RAND2, and RAND3).

	MFCC	VQual2	fusion
RAND1	0.295	0.300	0.563
RAND2	0.151	−0.099	0.220
RAND3	0.503	0.125	0.284

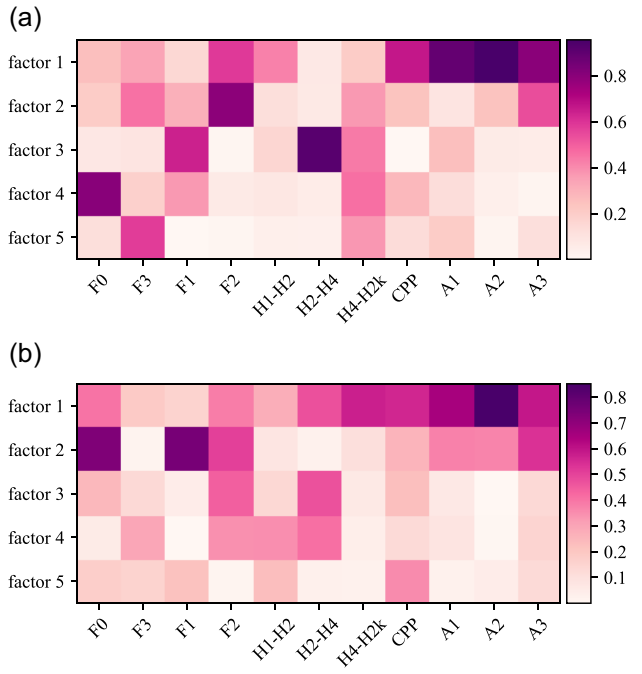

FIG. 2. (Color online) Absolute values of the factor loadings for acoustic measures. A five-dimensional factor analysis was performed using (a) the means and (b) standard deviations of the acoustic measures for each utterance for dimensionality reduction.

TABLE V. Multiple regression results on human and machine MDS coordinates (dependent variables) with acoustic speaker spaces (independent variables). The first three columns show R^2, F-statistics, and p-values of the multiple regression models. Only the MDS dimensions which can be modeled with $p < 0.05$ are shown in the table. SE, T, and p, which indicate the standard error, t-statistics, and p-values of the independent variables, are shown for each of the factors. The independent variables with $p < 0.05$ are boldfaced.

	Model			Factor 1			Factor 2			Factor 3			Factor 4			Factor 5		
	R^2	$F(5,9)$	p	SE	T	p	SE	T	p	SE	T	p	SE	T	p	SE	T	p
Means																		
RAND1 human D1	0.76	5.82	0.01	0.25	−0.48	0.64	0.31	−0.96	0.36	0.32	−2.05	0.07	**0.22**	**−2.74**	**0.02**	**0.15**	**−3.77**	**0.00**
RAND2 MFCC D2	0.74	5.13	0.02	0.41	2.17	0.06	**0.35**	**3.21**	**0.01**	0.28	−2.23	0.05	0.19	0.54	0.60	0.26	0.50	0.63
RAND2 fusion D1	0.70	4.18	0.03	0.42	2.22	0.05	0.36	1.49	0.17	0.29	1.10	0.30	**0.19**	**2.66**	**0.03**	0.27	−0.39	0.71
RAND2 human D1	0.86	11.44	0.00	0.35	−1.44	0.19	**0.30**	**−3.25**	**0.01**	0.24	−0.17	0.87	**0.16**	**−4.30**	**0.00**	0.22	−2.01	0.08
RAND3 human D1	0.90	15.66	0.00	0.19	−0.13	0.90	0.22	−1.93	0.09	0.15	0.88	0.40	0.15	−0.29	0.78	**0.13**	**−5.18**	**0.00**
Standard deviations																		
RAND1 MFCC D1	0.71	4.36	0.03	0.31	0.38	0.71	0.27	1.72	0.12	0.28	0.40	0.70	0.34	1.71	0.12	**0.35**	**3.23**	**0.01**
RAND2 MFCC D1	0.78	6.44	0.01	0.30	−0.20	0.85	0.25	−2.25	0.05	**0.28**	**−3.03**	**0.01**	**0.26**	**2.34**	**0.04**	0.26	−0.70	0.50
RAND3 MFCC D1	0.67	3.64	0.04	0.33	−0.59	0.57	0.32	−0.49	0.64	0.36	1.18	0.27	0.28	−1.37	0.20	**0.38**	**3.27**	**0.01**
RAND3 MFCC D2	0.92	19.81	0.00	**0.15**	**−5.32**	**0.00**	0.15	0.49	0.64	**0.17**	**−3.76**	**0.00**	0.13	−2.19	0.06	0.18	−1.74	0.12
RAND3 fusion D1	0.78	6.43	0.01	**0.30**	**2.39**	**0.04**	0.29	−0.73	0.48	0.33	1.37	0.20	0.26	1.57	0.15	0.34	−1.49	0.17
RAND3 human D1	0.66	3.51	0.05	**0.40**	**2.30**	**0.05**	0.39	0.20	0.84	0.44	0.63	0.54	0.34	0.72	0.49	0.46	−1.64	0.14

of voice quality, none of the MDS dimensions was significantly associated with any factor(s), even though the factors were estimated using voice quality features. R^2 values in linear regression only reflect the linear part of the decision-making process, but there are other parts that are not linear. Thus, even though the VQual2 system makes decisions based on the VQual2 feature set, its decision space might not be fully interpretable as a linear combination of features.

4. Discussion

Across all three subsets of the read-read, unmarked stimuli, humans were more accurate and consistent at speaker discrimination than were machines. However, subsets differed in how difficult they were for humans versus for machines, and human and machine speaker spaces were not strongly related in terms of the features that explained stimulus confusability. Differences between humans and machines could have occurred because humans utilized information that was not explicitly given to machines, such as spectro-temporal information and linguistic knowledge, and/or they used similar acoustic features but processed them differently.

The present results do not allow us to evaluate these possibilities. To evaluate the first possibility, an automatic system that can process supra-segmental information is needed. The widely used, frame-level feature based ASV system used in the current study is not explicitly given such information. Other systems that utilize prosodic information to model speakers (e.g., Dehak *et al.*, 2007) need to be developed. Evaluating the second possibility would require a more complex model of how features are processed and used in decision-making. For example, even though the VQual2 system made decisions based on voice quality features, the decisions did not appear to depend on the linear combination of the means and standard deviations of the features that explained humans' and MFCC system performance.

Instead, the results highlight differences in human and machine decision making. For example, the most important dimensions underlying human responses were highly related to the means of acoustic features, while MFCC responses were more closely related to standard deviations of the same features. This might indicate that humans perform best with the speakers whose speech varies widely in mean values, while MFCC-based systems work best when the within-utterance (voice quality) variance is large so that the acoustic information in an utterance is sufficient to model the speaker.

VI. CONCLUSION

Human and machine speaker discrimination performance on short-utterance, text-independent stimulus pairs were investigated in this study. Read sentences shorter than 2 s were used to evaluate performance with clear speech, and excerpts from pet-directed speech of similar duration were used to investigate the effect of exaggerated prosody. Analyses compared performance when pairs were matched (read-read) or mismatched (read-pet) for speaking style.

Results showed that human listeners were reasonably accurate at discriminating voices based on read-read pairs, but performance degraded significantly with style-mismatched pairs. Contrary to expectations, humans performed worse when discriminating between two marked speakers than when discriminating between two unmarked speakers, both for read-read pairs and read-pet pairs. The effect of speaker markedness on speaker discrimination is worth exploring in detail in the future. The UCLA Speaker Variability Database includes many non-native speakers of English whose speech could be useful for this purpose.

The machines tested here were less accurate than humans for read-read pairs, which is consistent with previous studies that reported poor ASV performance with short-utterance text-independent tasks. Performance degraded even more with pet-directed speech for unmarked speakers, especially for MFCC- and VQual2-based systems, either because prosody exaggeration distorted acoustic features or

because the databases used for the pre-training did not have a similar speaking style. Score-level fusion of the two systems improved performance, suggesting that VQual2 features provide information that is complementary to MFCCs. These features may be especially valuable when within-speaker variability is large. Interestingly, with style-mismatched pairs, speaker markedness had little effect on VQual2 features, and MFCC and fusion performance even improved for these pairs, to such an extent that machines outperformed human listeners. Unfortunately, the number of marked speakers in this study was not large enough to ensure that this result is robust. A follow-up study will analyze what advantage machines have when human performance is critically affected, and how to utilize that advantage in speaker verification tasks.

Human and machine performance on read-read pairs of unmarked speakers was further investigated with MDS on smaller subsets of speakers. CCA results between human and machine speaker spaces showed a weak relationship between the human and machine spaces. Further, better machine performance did not lead to an increase in the strength of this association. These results suggest that humans and machines use different strategies to distinguish speakers. Multiple regression between acoustic feature factors and MDS spaces for humans and machines found that human MDS axes were reasonably well-modeled as linear combinations of means of voice quality features. On the other hand, neither MFCC nor VQual2 MDS spaces could be well-modeled using mean values. These findings suggest that investigating how voice quality feature means are related to human responses might provide valuable insights into perceived speaker identity. Such knowledge could also prove useful for improving machine performance, by exploring how to process acoustic feature means effectively.

In future studies, we will examine human and machine performance differences in detail. Machine performance, for example, can be examined by varying the training data conditions, such as the speakers' language background, gender, and/or recording conditions. Modeling prosodic features and developing duration compensation techniques for very short utterances (2 s) might also be a promising research direction for ASV, as is examining how effectively human and machine decisions can be combined. Finally, using professional voice mimics or synthetic voices will allow for a more systematic evaluation of several acoustic factors.

ACKNOWLEDGMENTS

We wish to thank Dr. A. McCree and colleagues at the Johns Hopkins University Human Language Technology Center of Excellence for providing the i-vector/PLDA system and computational resources. This research was supported in part by NIH Grant No. DC01797 and by NSF Grant No. 1704167.

[1]*F*4, which has been shown to be important for discriminating male voices (Baumann and Belin, 2010; Nolan *et al.*, 2011), is not included in this feature set because the stimuli were band-limited at 4 kHz. However, this lack of *F*4 may be less problematic for discrimination of female voices, as in our study—Baumann and Belin found that for female voices, *F*1 was more important than higher formants.

[2]Automatic formant correction might be erroneous, especially when formant frequencies are close to harmonics, resulting in overcorrection. It is also possible that uncorrected harmonic amplitudes contain speaker-specific information that is useful for ASV.

[3]We tried to match the utterance duration between the training data and evaluation data for i-vector and PLDA training by truncating the original recordings to 2-s segments. However, it did not show any notable performance differences, possibly due to decreased phonetic coverage. Thus, the original recordings were retained.

Baumann, O., and Belin, P. (**2010**). "Perceptual scaling of voice identity: Common dimensions for different vowels and speakers," Psychol. Res. **74**(1), 110–120.

Bricker, P. D., and Pruzansky, S. (**1966**). "Effects of stimulus content and duration on talker identification," J. Acoust. Soc. Am. **40**, 1441–1449.

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (**2002**). "What's new, pussycat? On talking to babies and animals," Science **296**(5572), 1435.

Cook, S., and Wilding, J. (**1997**). "Earwitness testimony: Never mind the variety, hear the length," Appl. Cogn. Psychol. **11**(2), 95–111.

Das, R. K., Jelil, S., and Prasanna, S. R. M. (**2016**). "Significance of constraining text in limited data text-independent speaker verification," in *Proceedings of the International Conference on Signal Processing and Communications (SPCOM)*, April 6–8, Bangalore, India, pp. 1–5.

Das, R. K., and Prasanna, S. R. M. (**2016**). "Exploring different attributes of source information for speaker verification with limited test data," J. Acoust. Soc. Am. **140**(1), 184–190.

Das, R. K., and Prasanna, S. R. M. (**2017**). "Speaker verification from short utterance perspective: A review," IETE Tech. Rev. 1–19.

Dehak, N., Dumouchel, P., and Kenny, P. (**2007**). "Modeling prosodic features with joint factor analysis for speaker verification," IEEE Trans. Audio Speech Lang. Process. **15**(7), 2095–2103.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (**2011**). "Front-end factor analysis for speaker verification," IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798.

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.

Goldstein, A. G., Knight, P., Bailis, K., and Conover, J. (**1981**). "Recognition memory for accented and unaccented voices," Bull. Psychonom. Soc. **17**(5), 217–220.

Hansen, J. H. L., and Hasan, T. (**2015**). "Speaker recognition by machines and humans: A tutorial review," IEEE Signal Process. Mag. **32**(6), 74–99.

Hanson, H. M. (**1997**). "Glottal characteristics of female speakers: Acoustic correlates," J. Acoust. Soc. Am. **101**(1), 466–481.

Hautamäki, V., Kinnunen, T., Nosratighods, M., Lee, K.-A., Ma, B., and Li, H. (**2010**). "Approaching human listener accuracy with modern speaker verification," in *Proceedings of Interspeech*, September 26–30, Makuhari, Chiba, Japan, pp. 1473–1476.

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 practices for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**(3), 225–246.

Iseli, M., Shue, Y.-L., and Alwan, A. (**2007**). "Age, sex, and vowel dependencies of acoustic measures related to the voice source," J. Acoust. Soc. Am. **121**(4), 2283–2295.

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

Kahn, J., Audibert, N., Rossato, S., and Bonastre, J. F. (**2011**). "Speaker verification by inexperienced and experienced listeners vs. speaker verification system," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 22–27, Prague, Czech Republic, pp. 5912–5915.

Keating, P. A., Kreiman, J., and Alwan, A. (**2018**). "The UCLA speaker variability database," https://ucla.box.com/s/81ho14uypkmv2nn1s3tvy1gvdtq70gcf (Last viewed July 6, 2018).

Kenny, P., Stafylakis, T., Ouellet, P., Alam, M. J., and Dumouchel, P. (**2013**). "PLDA for speaker verification with utterances of arbitrary duration," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 26–31, Vancouver, Canada, pp. 7649–7653.

Kreiman, J., and Gerratt, B. R. (**1996**). "The perceptual structure of pathologic voice quality," J. Acoust. Soc. Am. **100**(3), 1787–1795.
Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., and Zhang, Z. (**2014**). "Toward a unified theory of voice production and perception," Loquens **1**(1), 1–9.
Kreiman, J., and Papcun, G. (**1991**). "Comparing discrimination and recognition of unfamiliar voices," Speech Commun. **10**(3), 265–275.
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 of Voice Studies* (Wiley-Blackwell, Oxford, UK).
Kruskal, J. B., and Wish, M. (**1978**). *Multidimensional Scaling* (Sage, Beverly Hills, CA).
Larcher, A., Lee, K. A., Ma, B., and Li, H. (**2014**). "Text-dependent speaker verification: Classifiers, databases and RSR2015," Speech Commun. **60**, 56–77.
Legge, G. E., Grosmann, C., and Pieper, C. M. (**1984**). "Learning unfamiliar voices," J. Exp. Psychol. Learn. Mem. Cogn. **10**(2), 298–303.
Lu, X., and Dang, J. (**2008**). "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," Speech Commun. **50**(4), 312–322.
Macmillan, N. A., and Creelman, C. D. (**2005**). *Detection Theory: A Users Guide*, 2nd ed. (Erlbaum, Mahwah, NJ).
Martin, A. F., and Greenberg, C. S. (**2009**). "NIST 2008 speaker recognition evaluation: Performance across telephone and room microphone channels," in *Proceedings of Interspeech*, September 6–10, Brighton, UK, pp. 2579–2582.
McDougall, K. (**2013**). "Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades," Int. J. Speech Lang. Law **20**(2), 163–172.
Nakasone, H., and Beck, S. D. (**2001**). "Forensic automatic speaker recognition," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, June 18–22, Crete, Greece, pp. 1–6.
Nicholson, S., Milner, B., and Cox, S. (**1997**). "Evaluating feature set performance using the F-ratio and J-measures," in *Proceedings of Eurospeech*, September 22–25, Rhodes, Greece, pp. 413–416.
Nolan, F., McDougall, K., and Hudson, T. (**2011**). "Some acoustic correlates of perceived (Dis) similarity between same-accent voices," in *Proceedings of ICPhS*, August 17–21, Hong Kong, China, pp. 1506–1509.
Park, S. J., Sigouin, C., Kreiman, J., Keating, P. A., Guo, J., Yeung, G., Kuo, F.-Y., and Alwan, A. (**2016**). "Speaker identity and voice quality: Modeling human responses and automatic speaker recognition," in *Proceedings of Interspeech*, September 8–12, San Francisco, CA, pp. 1044–1048.
Park, S. J., Yeung, G., Kreiman, J., Keating, P. A., and Alwan, A. (**2017**). "Using voice quality features to improve short-utterance, text-independent speaker verification systems," in *Proceedings of Interspeech*, August 20–24, Stockholm, Sweden, pp. 1522–1526.
Parthasarathy, S., Zhang, C., Hansen, J. H. L., and Busso, C. (**2017**). "A study of speaker verification performance with expressive speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 5–9, New Orleans, LA, pp. 5540–5544.
Przybocki, M., and Martin, A. (**2004**). "NIST speaker recognition evaluation chronicles," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, May 31–June 3, Toledo, Spain, pp. 12–22.
Przybocki, M., Martin, A., and Le, A. (**2006**). "NIST speaker recognition evaluation chronicles—Part 2," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, June 28–30, San Juan, Puerto Rico, pp. 1–6.
Ramachandran, R. P., Farrell, K. R., Ramachandran, R., and Mammone, R. J. (**2002**). "Speaker recognition-general classifier approaches and data fusion methods," Pattern Recogn. **35**(12), 2801–2821.
Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Qin Jin, Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., and Xiang, B. (**2003**). "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 6–10, Hong Kong, China, pp. IV–784-7.
Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (**2000**). "Speaker verification using adapted gaussian mixture models," Digital Signal Process. **10**(1–3), 19–41.
Roebuck, R., and Wilding, J. (**1993**). "Effects of vowel variety and sample length on identification of a speaker in a line-up," Appl. Cogn. Psychol. **7**(6), 475–481.
Saslove, H., and Yarmey, A. D. (**1980**). "Long-term auditory memory: Speaker identification," J. Appl. Psychol. **65**(1), 111–116.
Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., and Zäske, R. (**2014**). "Speaker perception," Cogn. Sci. **5**(1), 15–25.
Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., and Stolcke, A. (**2005**). "Modeling prosodic feature sequences for speaker recognition," Speech Commun. **46**(3–4), 455–472.
SYSTAT (**2018**). "Version 13.2," Systat Software Inc., San Hose, CA.
Tabachnick, B. G., and Fidell, L. S. (**2013**). *Using Multivariate Statistics*, 6th ed. (Pearson, New York).
Van Lancker, D., and Kreiman, J. (**1987**). "Voice discrimination and recognition are separate abilities," Neuropsychologia **25**(5), 829–834.
Vijayan, K., Raghavendra Reddy, P., and Sri Rama Murty, K. (**2016**). "Significance of analytic phase of speech signals in speaker verification," Speech Commun. **81**, 54–71.