INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition

Soo Jin Park[1], *Caroline Sigouin*[2], *Jody Kreiman*[3], *Patricia Keating*[4], *Jinxi Guo*[1], *Gary Yeung*[1], *Fang-Yu Kuo*[1] *and Abeer Alwan*[1]

[1]Department of Electrical Engineering, University of California, Los Angeles, USA
[2]Department of Languages, Linguistics and Translation, Université Laval, Québec, Canada
[3]Dept of Head and Neck Surgery, School of Medicine, University of California, Los Angeles, USA
[4]Department of Linguistics, University of California, Los Angeles, USA

sj.park@ucla.edu, caroline.sigouin.1@ulaval.ca, jkreiman@ucla.edu, keating@humnet.ucla.edu
lennyguo@g.ucla.edu, garyyeung@g.ucla.edu, fy.kuo@ucla.edu, alwan@ee.ucla.edu

Abstract

Despite recent breakthroughs in automatic speaker recognition (ASpR), system performance still degrades when utterances are short and/or when within-speaker variability is large. This study used short test utterances (2-3sec) to investigate the effect of within-speaker variability on state-of-the-art ASpR system performance. A subset of a newly-developed UCLA database is used, which contains multiple speech tasks per speaker. The short utterances combined with a speaking-style mismatch between read sentences and spontaneous affective speech degraded system performance, for 25 female speakers, by 36%. Because humans are more robust to utterance length or within-speaker variability, understanding human perception might benefit ASpR systems. Perception experiments were conducted with recorded read sentences from 3 female speakers, and a model is proposed to predict the perceptual dissimilarity between tokens. Results showed that a set of voice quality features including F0, F1, F2, F3, H1*-H2*, H2*-H4*, H4*-H2k*, H2k*-H5k, and CPP provides information that complements MFCCs. By fusing the feature set with MFCCs, human response prediction RMS error was .12, which represents a 12% relative error reduction compared to using MFCCs alone. In ASpR experiments with short utterances from 50 speakers, the voice quality feature set decreased the error rate by 11% when fused with MFCCs.

Index Terms: voice quality, speech perception model, speaker recognition

1. Introduction

Recognizing the identity of a speaker from his or her voice is an important topic for researchers in various fields including forensic phonetics, voice therapy, and engineering. In automatic speaker recognition (ASpR), the emergence of the i-vector framework brought about remarkable improvements. Machines could outperform humans in some difficult conditions (e.g. channel mismatch, non-native language, unfamiliar voices) [1]. However, machines require a large amount of enrollment and test data and long utterances to model speakers' voices. For example, when the enrollment and test utterance lengths are shortened to 2 sec from 2.5 min, the equal error rate (EER) surges to 35% from 3.37% for the NIST 2008 SRE telephone-based utterances [2]. Performance also degrades when within-speaker variability is large, as in emotional speech [3]. Humans, on the other hand, are able to distinguish speakers with high accuracy even from very short utterances. If the listeners are familiar with the speaker, humans perform better with short utterances and emotional variability than machines [4]. Thus, obtaining insights into how humans recognize speakers may improve ASpR system performance. Modeling human responses to predict perceived speaker identity itself is an interesting topic. In forensics, for instance, measuring the similarity between speakers is a critical issue in constructing fair voice line-ups for ear witnesses [5].

Finding features representing speaker-specific information is critical both for human response modeling and for ASpR. Yet, no single set of acoustic parameters associated with human speaker recognition has been identified [6]. Studies have shown that humans recognize familiar voices as complex, integral auditory patterns [6, 7, 8], and that they remember unfamiliar voices with reference to a "prototype" and deviations from that prototype [6, 9]. In ASpR systems, the most popular features are mel-frequency cepstral coefficients (MFCCs) [10]. Note that they represent vocal tract information well, but not the voice source. Although the voice source contains abundant speaker-specific information, only a few studies have used it in ASpR applications. For example, Espy-Wilson et al. used eight acoustic parameters consisting of both voice source and vocal tract features [11]. Mazaira-Fernandez et al. separated voice source from vocal tract information, and they combined cepstral coefficients from those two estimates [12]. These studies showed the effectiveness of voice source information with promising results, but such information still has not been utilized extensively in ASpR.

We used voice quality features in order to reflect voice source characteristics. Here, the term *voice quality* is defined as a perceptual response to an acoustic voice signal, and quality is measured using a psychoacoustic model proposed in [13]. This model includes perceptually-validated spectral domain parameters to model the voice source, along with formant frequencies and additional parameters to model F0, amplitude, and the inharmonic part of the voice source. In a previous study, these acoustic voice quality measures predicted listeners' confusions between 5 female speakers reasonably well from sustained vowel /a/ sounds [14]. This study attempted to use this same set of acoustic features to predict perceptual dissimilarity from sentences read by 3 speakers, as well as vowel sounds. If features from the psychoacoustic model are useful to predict

http://dx.doi.org/10.21437/Interspeech.2016-523

perceptual dissimilarity, it is possible that these features will also prove useful in ASpR.

In this study, we aimed to answer the following questions: How well do humans recognize speaker identity, and are voice quality features useful to model human responses? How can voice quality features be used to improve ASpR systems when there is variability and the utterances are short?

2. Data and Acoustic Analysis

2.1. Database

A database to study within- and between-speaker variability was designed and collected at UCLA. Over two hundred female and male speakers participated in three separate recording sessions on different days. At the beginning of each session, they repeated the isolated vowel /a/ three times and read two repetitions of five Harvard sentences. These tasks allow cross-session comparison.

Two further speech tasks were included in each session. They were designed to sample variability consistent with normal everyday speaking situations. For example, to collect speech samples with affect variability, speakers were asked to report on a recent conversation which was neutral, happy, or annoying. Other speech tasks included giving instructions, making a phone call, and speaking to pets. All of the audio recordings were made in a sound-attenuated booth with a sampling rate of 22 kHz. Detailed description of the database can be found in [14].

2.2. Voice Quality Features

A voice quality feature set was used to measure the correlation between acoustic and perceived within-speaker variability in our previous study [14]. The set consists of F0, F1, F2, F3, H1*-H2* (the amplitude difference between the first and second harmonics), H2*-H4* (the amplitude difference between the second and the forth harmonic), the difference between H4* and H2k* (the amplitude of harmonic component near 2 kHz), the difference between H2k* and H5k, and cepstral peak prominence (CPP, [15]). The asterisks (*) indicate that the effect of formants on harmonic amplitudes is corrected [16]. These features are selected based on an extensive study performed to find the necessary and sufficient set of features contributing to perceived voice quality [13, 17, 18, 19, 20]. This study applied these features to predict perceived dissimilarity between voices, and to improve automatic speaker recognition systems.

The measures were extracted with VoiceSauce software [21] every 10 msec using Straight [22] for pitch estimation and Praat [23] for extracting formant frequencies. The window length was set to 25 msec for formant extraction, but the source measures were extracted pitch-synchronously.

3. Perceptual Experiments and Modeling

3.1. Perceptual Experiments

Eight read sentences recorded from 3 female speakers were used as stimuli in the experiment. Two sessions per speaker were selected, with 2 repetitions of 2 different sentences per session. The selected sentences were "A pot of tea helps to pass the evening" and "The soft cushion broke the man's fall". The stimuli pairs were used to compare the effect of session, content, and speaker difference. From a total of 24 tokens, 30 same-speaker pairs and 48 different-speaker pairs were created.

To determine how easy these different sentences were to distinguish, 15 normal-hearing UCLA students and staff members participated in a listening experiment. They heard 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). Listeners were not told how many speakers were represented in the trials. They judged whether the voices represent one speaker or two different speakers for each pair of stimuli, and reported their confidence in their response on a scale from 1 (positive) to 5 (wild guess). The experiment was self-paced and listeners were encouraged to take breaks as needed.

Similar experiments were conducted with 60 listeners and sustained vowel /a/ sounds from 5 female speakers, as described in [14].

3.2. Human Listener Performance

Human listeners were quite accurate in distinguishing same/different speaker pairs even when the utterances were short ($\leq$ 3 sec). For these sentence stimuli, listeners averaged 89.0% correct (hits & correct rejections) (sd: 8.21%, range: 65.4%-97.4%). In comparison, our previous study using isolated vowels reported accuracy ranging from 38.6% to 84.8%, with a mean of 69.0% (sd: 10.43%) [14].

The dissimilarity score d from an individual listener was calculated from the listener's same/different speaker response and the uncertainty u which was reported on a 1 (positive) to 5 (wild guess) scale. If a listener responded that he/she was positive the two tokens were from the same speaker ($u = 1$), then d should be low. On the other hand, if the response was "positive these are different speakers", then d should be high. In this sense, the dissimilarity score was defined in following way:

$$d = \begin{cases} u & \text{if "same speaker" response} \\ 11 - u & \text{if "different speaker" response} \end{cases}$$

The dissimilarity score was then averaged across listeners. The resulting averaged dissimilarities $\bar{d}$ ranged from 0 to 10, where '0' was assigned to identical token pairs, which were not included in the perception experiment.

It was observed that when the same/different speaker decision was re-calculated based on the ensemble score $\bar{d}$, accuracy increased substantially to 97.4% (see Figure 1), versus 89.0% for averaged data. The accuracy gain is more obvious in the vowel case, for which ensemble accuracy reached 86.4%, which is higher than the score of the best individual listener.

3.3. Modeling Human Responses

Multi-dimensional scaling (MDS, [24]) was used to compute the distance between tokens in a perceptual space. The averaged dissimilarity score $\bar{d}$ was normalized to have a value between 0 and 1, and the normalized score was analyzed using a 6-dimensional non-metric MDS (stress=0.004 and 0.058 for sentences and vowels). The Euclidean distances between token pairs of all possible combinations were calculated in the MDS space. The objective of using MDS was to obtain perceptual distance, not to reduce dimensionality for visualization. Therefore, higher dimensional MDS was used to represent the dissimilarity between stimuli [25, 26]. The resulting token distance had a 0 to 1 range.

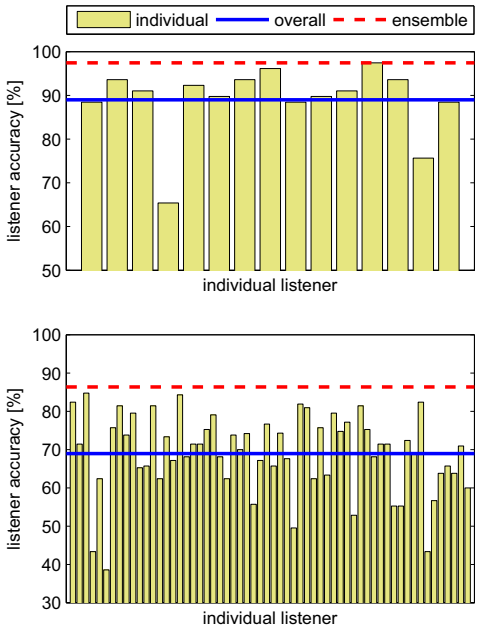

Figure 1: *Listener performance (accuracy) in identifying same/different speaker pairs with read sentences and 15 listeners (top), and sustained vowel /a/ sounds and 60 listeners (bottom).*

A standard feature set for ASpR systems was used as a baseline to predict human responses. This set included 20-MFCCs along with their first- and second-order derivatives. Voice quality (VQual) features described in Section 2.2 with their first- and second-order derivatives were also included. The mean and standard deviation of each feature including derivatives within a token were calculated, and the absolute differences in the feature mean and standard deviation were found between tokens. The perceptual dissimilarities were predicted with a linear regression framework. Here, the variable being predicted was the Euclidean distance between two tokens in the MDS perceptual space, and the predictors were the differences in means and standard deviations between the two tokens. The predictors were of 120-dim for MFCCs and 54-dim for VQuals. MFCCs and VQuals were used individually or combined by concatenating them together before linear regression.

3.4. Results

The human response prediction results in terms of root-mean-squared error (RMSE) between the predicted value and the token distance in the MDS space, either with only the mean or with the mean and standard deviation of every feature, are summarized in Table 1. As hypothesized, VQuals provided complementary information to MFCCs. When the perceptual distance was predicated with MFCC feature vectors, consisting of the mean and standard deviation of each feature and its derivatives, the best RMSE was 0.140 for sentences and 0.121 for vowels. Using only VQual features did not improve the performance. However, when they were combined with MFCCs, the RMSE performance improved by 11.80% for sentences.

Combined MFCC and VQual means improved relative performance by only 2.24%, while combined mean and standard deviation (mean&sd) improved 11.80% for sentences. On the other hand, mean&sd for vowels improved the performance only by 3.14%, possibly because sustained vowel sounds do not vary much within a token.

Note that the human listeners had higher accuracy on read sentences than on isolated vowels, but the acoustic features did less well at predicting human performance for the sentences. This was expected because there are many other sources of information in connected speech that are not represented by the current feature set.

Score-level fusion was also tried, but no improvement was found over concatenating the features.

Table 1: *Perceptual dissimilarity prediction performance in RMSE using either MFCC, VQual, and the combination of the two. Relative improvements by combining MFCC and VQual features compared to the performance with only MFCC are shown in parentheses.*

	sentences		vowels	
	mean	mean&sd	mean	mean&sd
MFCC	0.143	0.140	0.123	0.121
VQual	0.156	0.140	0.128	0.128
MFCC+VQual	0.140 (2.24%)	0.123 (11.80%)	0.118 (4.07%)	0.117 (3.14%)

In summary, human response prediction showed the effectiveness of the voice quality feature set. Before applying these features to ASpR systems, we first examined how a state-of-the-art ASpR system would perform with short utterances and within-speaker variability.

4. Automatic Speaker Recognition

4.1. Standard ASpR System Setup

The effect of within-speaker variability can be observed by comparing results from two different conditions. One is to enroll the speakers with data containing variability and test with known variability (*matched condition*), and the other is to test with unseen variability (*mismatched condition*). For example, in the session-matched condition, each speaker is enrolled with randomly selected samples from all three sessions and tested on the remaining tokens. In the session-mismatched condition, the speakers are enrolled using only data from two sessions and tested on the third. Affect- and style-matched and mismatched conditions are defined in a similar way.

The types of within-speaker variability of interest in this study are session, affect, and speaking-style variability. In the UCLA database, read sentences in all 3 sessions and spontaneous speech with differing affect at each session are available. These speech samples were randomly selected for 25 female and 25 male speakers. Because automatic speaker recognition (ASpR) systems are sensitive to the utterance length of the enrollment and test data, it is important to balance the amount of data for a fair comparison. In each session per speaker, there are ten read sentences, each is 2-3 sec long, and an affective speech recording lasting 30-60 sec. In order to balance the amount of data between read sentences and affective speech, we clipped a 30-sec segment in the middle of the affective speech and divided the segment into ten 3-sec segments. The resulting amount of data to enroll each speaker was approximately 60 sec for session and affect variability experiments, and approximately 90 sec for style variability experiments. All test utterances were 2-3 sec long.

The ASpR system performance was evaluated with a state-of-the-art system. The 20-order MFCCs with their first- and second-order derivatives were used as features. An i-vector [27]/PLDA [28] speaker verification system was implemented with the Kaldi toolkit [29]. The system was developed in a gender-dependent way.

4.2. Standard ASpR System Performance

The ASpR system performance with short test utterances, influenced by within-speaker variability, is reported in terms of equal error rates (EER) in Table 2. Session variability did not affect system performance and hence, those results are not shown in the table.

Affect variability and speaking-style variability, however, caused a notable degradation in system performance, both for female and male speakers. Affect variability among neutral, happy, and annoyed speech by female speakers caused the most degradation, doubling the error rate. Note that the affective speech recordings in the database also had session variability because they were recorded in different sessions. However, since the effect of the session variability was negligible, as mentioned earlier, the affect variability was regarded as the main reason for the performance degradation.

It is possible that these results are dependent on the lexical content. The read sentences were distributed evenly into enrollment and test sets so that the system was enrolled with all 5 different sentences. For affective speech, however, lexical imbalance could occur between enrollment and test data because lexical content was not controlled. Further analysis is needed with more speakers and more controlled conditions.

Table 2: *Equal error rate (EER) for the ASpR system, using only MFCC features, for the different conditions. The relative error increase in the mismatched compared to the matched conditions is reported in parentheses.*

	female	male
affect-matched	3.64%	2.67%
affect-mismatched	7.49% (105.68%)	4.00% (50.00%)
style-matched	5.07%	2.37%
style-mismatched	6.87% (35.64%)	3.64% (53.91%)

4.3. Voice Quality Feature Effect on the ASpR System

The standard system described in Section 4.1 was used in this analysis. As baseline features, 20-MFCCs with their first- and second-derivatives were applied to the system. Another system with the same back-end but with voice quality features was also implemented. Because the development data (NIST SRE) were sampled at 8kHz, the measure H2k*-H5k could not be used, since it requires access to the harmonic component close to 5kHz. Thus, this feature was excluded from the VQual set in this task. The resulting feature vector dimension was 60 for MFCCs and 24 for VQuals. The two systems were fused at the score level to obtain final results.

It was found that the voice quality features provided complementary information to MFCCs. The performance of the ASpR system which fused MFCCs and VQuals is summarized in Table 3. In the affect-matched condition, fusing voice quality features improved the system performance for both genders. The improvement was 11.60% for female voices and 19.11% for male voices. A similar trend was observed for the affect-mismatched cases and style-matched conditions. However, the style-mismatched condition degraded slightly for female speakers when adding VQual features.

Even though the EER improved by adding VQual features, a difference in EER between matched and mismatched conditions remains. This can be explained by the fact that voice quality may be varying significantly according to the emotional status and speaking-style of the speaker. Further analysis is needed with orthographic transcriptions and acoustic measures. Nevertheless, it was apparent that voice quality features provided speaker-specific information which might not be sufficiently represented by MFCCs.

Table 3: *Equal error rate (EER) for the proposed system. The relative improvements over using only MFCCs are shown in parentheses.*

	female	male
affect-matched	3.22% (11.60%)	2.16% (19.11%)
affect-mismatched	6.70% (10.49%)	3.60% (9.89%)
style-matched	4.65% (8.26%)	2.16% (8.78%)
style-mismatched	6.90% (−0.37%)	3.38% (7.25%)

5. Conclusion

This study investigated the effectiveness of voice quality features (VQual) based on a psychoacoustic model. These features were used to model human responses in a series of perceptual experiments. In predicting perceived speaker dissimilarities, VQuals provided complementary information to MFCCs. The root-mean-squared error decreased as much as 11.80% for read sentences by combining the means and standard deviations of VQuals and MFCCs.

VQual and MFCC features were then applied to an automatic speaker recognition (ASpR) system. The experiments were conducted with a newly developed database containing various kinds of within-speaker variability. It was found that a state-of-the-art ASpR system performed worse when there was within-speaker variability and utterances were short ($\leq$ 3 sec). VQuals did not necessarily improve the robustness to within-speaker variability, but they did improve ASpR system performance in most conditions. However, the reliability of VQual features highly depends on pitch tracking and formant tracking performance. This might make it hard to apply this set of features to utterances with extremely high/low pitch.

Further studies will include perception experiments with more speakers and more within-speaker variability to reveal how robust humans are to a wide range of variabilities. The knowledge gained might also improve ASpR system robustness to such variability. Higher level features such as prosodic features will also be considered for further improvements.

6. Acknowledgements

This research was supported in part by NSF and NIH grant DC01797.

7. References

[1] J. Kahn, N. Audibert, S. Rossato, and J. F. Bonastre, "Speaker verification by inexperienced and experienced listeners vs. speaker verification system," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 5912–5915.

[2] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "i-vector based speaker recognition on short utterances," in *Interspeech*, 2011, pp. 2341–2344.

[3] T. Wu, Y. Yang, Z. Wu, and D. Li, "MASC: A speech corpus in mandarin for emotion analysis and affective speaker recognition," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, June 2006, pp. 1–5.

[4] S. J. Wenndt and R. L. Mitchell, "Machine recognition vs human recognition of voices," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 4245–4248.

[5] F. Nolan, K. Mcdougall, and T. Hudson, "Some acoustic correlates of perceived (dis) similarity between same-accent voices," in *International Congress of Phonetic Sciences (ICPhS)*, 2011, pp. 1506–1509.

[6] J. Kreiman and D. Sidtis, *Foundations of voice studies*. Wiley-Blackwell, 2011.

[7] D. Van Lancker, J. Kreiman, and T. D. Wickens, "Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices." *Journal of Phonetics*, vol. 13, no. 1, pp. 39–52, 1985.

[8] D. Van Lancker and J. Kreiman, "Voice discrimination and recognition are separate abilities," *Neuropsychologia*, vol. 25, no. 5, pp. 829–834, 1987.

[9] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Communication*, vol. 10, no. 3, pp. 265–275, 1991.

[10] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, Nov 2015.

[11] C. Y. Espy-wilson, E. Manocha, and S. Vishnubhotla, "A new set of features for textindependent speaker identification," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2006, pp. 1475–1478.

[12] L. M. Mazaira-Fernandez, A. Álvarez-Marquina, and P. Gómez-Vilda, "Improving speaker recognition by biometric voice deconstruction," *Frontiers in bioengineering and biotechnology*, vol. 3, 2015.

[13] M. Garellek, R. Samlan, B. Gerratt, and J. Kreiman, "Modeling the voice source in terms of spectral slopes," *J. Acoust. Soc. Am.*, vol. 139, pp. 1404–1410, 2016.

[14] J. Kreiman, S. J. Park, P. A. Keating, and A. Alwan, "The relationship between acoustic and perceived intraspeaker variability in voice quality," in *Interspeech*, 2015, pp. 2357–2360.

[15] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *J. Speech Lang. Hear. Res.*, vol. 37, no. 4, pp. 769–778, 1994.

[16] M. Iseli, Y.-L. Shue, and A. Alwan, "Age, sex, and vowel dependencies of acoustic measures related to the voice source," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2283–2295, 2007.

[17] J. Kreiman and B. R. Gerratt, "Perceptual sensitivity to first harmonic amplitude in the voice source," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2085–2089, 2010.

[18] ——, "Perceptual interaction of the harmonic source and noise in voice," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 492–500, 2012.

[19] M. Garellek, P. Keating, C. M. Esposito, and J. Kreiman, "Voice quality and tone identification in White Hmong," *J. Acoust. Soc. Am.*, vol. 133, no. 2, pp. 1078–1089, 2013.

[20] M. Garellek, R. A. Samlan, J. Kreiman, and B. R. Gerratt, "Perceptual sensitivity to a model of the source spectrum," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, 2013.

[21] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *International Congress of Phonetic Sciences (ICPhS)*, 2011, pp. 1846–1849.

[22] H. Kawahara, A. de Cheveigné, and R. D. Patterson, "An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: Revised TEMPO in the STRAIGHT-suite." in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 1998.

[23] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[24] J. B. Kruskal and M. Wish, *Multidimensional scaling*. Sage, 1978, vol. 11.

[25] J. Jiang, E. T. Auer, A. Alwan, P. A. Keating, and L. E. Bernstein, "Similarity structure in visual speech perception and optical phonetic signals," *Perception & Psychophysics*, vol. 69, no. 7, pp. 1070–1083, 2007.

[26] K. McDougall, F. Nolan, and T. Hudson, "Telephone transmission and earwitnesses: Performance on voice parades controlled for voice similarity," *Phonetica*, vol. 72, no. 4, pp. 257–272, 2015.

[27] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

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

[29] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The Kaldi speech recognition toolkit," in *In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.