INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Using Voice Quality Features to Improve Short-Utterance, Text-Independent Speaker Verification Systems

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

[1]Dept of Electrical Engineering, University of California Los Angeles, USA
[2]Dept of Head and Neck Surgery, School of Medicine, University of California Los Angeles, USA
[3]Dept of Linguistics, University of California Los Angeles, USA

sj.park@ucla.edu, garyyeung@g.ucla.edu, jkreiman@ucla.edu, keating@humnet.ucla.edu, alwan@ee.ucla.edu

Abstract

Due to within-speaker variability in phonetic content and/or speaking style, the performance of automatic speaker verification (ASV) systems degrades especially when the enrollment and test utterances are short. This study examines how different types of variability influence performance of ASV systems. Speech samples ($<$ 2 sec) from the UCLA Speaker Variability Database containing 5 different read sentences by 200 speakers were used to study content variability. Other samples (about 5 sec) that contained speech directed towards pets, characterized by exaggerated prosody, were used to analyze style variability. Using the i-vector/PLDA framework, the ASV system error rate with MFCCs had a relative increase of at least 265% and 730% in content-mismatched and style-mismatched trials, respectively. A set of features that represents voice quality (F0, F1, F2, F3, H1-H2, H2-H4, H4-H2k, A1, A2, A3, and CPP) was also used. Using score fusion with MFCCs, all conditions saw decreases in error rates. In addition, using the NIST SRE10 database, score fusion provided relative improvements of 11.78% for 5-second utterances, 12.41% for 10-second utterances, and a small improvement for long utterances (about 5 min). These results suggest that voice quality features can improve short-utterance text-independent ASV system performance.

Index Terms: speaker recognition, within-speaker variability, voice quality

1. Introduction

A single speaker's voice can vary dramatically in different situations. Word choices, mood, intentions, health conditions, and the relationship to the listener all affect the acoustic characteristics of that person's voice. Such within-speaker variability causes major difficulties when identifying speakers from their voices. This problem becomes critical when the utterances used to enroll and verify speakers are short. For instance, the equal error rate (EER) for text-independent automatic speaker verification (ASV) is 1.59–2.48% for 2-minute utterances, while the EER skyrockets to 10.52–21.83% for 5-second utterances [1, 2]. A possible interpretation of this phenomenon is that shorter utterances cannot capture all the variability in a speaker's voice. This within-speaker variability falls into two categories: extrinsic variability and intrinsic variability [3]. Extrinsic variability includes variability that is out of the speaker's control, such as recording conditions, channel types, and noise. Intrinsic variability includes variability that characterizes the speaker's voice, such as word choice, articulation, emotion, and speaking style. Although extrinsic variability also affects the system performance, we focus on intrinsic variability in this study. We are most interested in finding speaker-characterizing features that are robust to the intrinsic variability, even in short utterances.

Conventional acoustic features such as mel-frequency cepstral coefficients (MFCCs) are effective in various speech processing applications, but they might not be sufficient for ASV when within-speaker variability is large. For instance, while MFCCs are successful at capturing the overall spectral envelope, they obscure fine vocal structures, which also have an abundance of speaker-specific information. Because the spectral envelope varies based on phonetic content, long segments of speech with rich phonetic content perform well with MFCCs, but shorter speech segments usually lack the variety of content needed. Researchers have attempted to find alternative features. For example, it has been found that voice source-related features improve speaker recognition systems by providing information that complements conventional cepstral features [4, 5, 6]. Das et al. also reported that features extracted from the voice source signal outperform MFCCs in ASV with test utterances shorter than 3 seconds [7]. In this study, various *voice quality* features are investigated.

Voice quality can be thought of as the "timbre of the voice". Although it is often associated with the voice source characteristics, vocal tract characteristics are also reflected. Laver, in his pivotal study, defined voice quality as the characteristic auditory coloring of an individual speaker's voice, encompassing both laryngeal and supra-laryngeal features [8]. Voice quality has recently gained momentum in speaker recognition communities because humans utilize voice quality to recognize speakers [9, 10]. Even though machines outperform humans in some long-utterance tasks [3, 11], human listener performance does not degrade much when the phonetic content and utterance lengths are limited [12]. These findings suggest that voice quality might provide important information for short-utterance text-independent ASV.

In previous work, we have shown that a voice quality feature set inspired by a psycho-acoustic model can predict human speaker perception and improve ASV performance by providing complementary information to MFCCs [13, 14]. In the present study, we extend our previous work by analyzing two types of within-speaker variability: phonetic content and speaking-style variability. Specifically, we address the following questions: 1) Which voice quality features are able to separate speakers when there is large within-speaker variability? 2) How much does the performance of a state-of-the-art ASV system degrade from content/style variability when the utterances are short, and how much help can the voice quality features contribute in such cases? 3) Would the voice quality features be useful for general short-utterance ASV tasks?

http://dx.doi.org/10.21437/Interspeech.2017-157

The remaining paper is organized as follows. Section 2 describes the database and the process used to choose features. Section 3 discusses the experiments and the results, comparing the content/style matched and mismatched trials. Section 4 evaluates the proposed feature set on a standard speaker recognition evaluation database. Section 5 concludes the paper with a brief summary and description of future work.

2. Data and Feature Selection

2.1. Database

The UCLA Speaker Variability Database [13] was developed to study both within- and between-speaker variability. Speech samples from more than 100 female and 100 male undergraduate students speaking in a variety of styles were collected across 3 different recording sessions per speaker. The speaking styles included read sentences; giving instructions; affective recountings of neutral, annoyed, and happy conversations; a phone-call; and pet-directed speech. All the audio recordings were made in a sound-attenuated booth with a sampling rate of 22 kHz. Among the various speaking styles, the read sentences and pet-directed speech were used in this study. Read sentences were used to analyze content variability, and both read sentences and pet-directed speech were used to analyze style variability.

The read sentences consisted of 2 repetitions of 5 Harvard sentences [15], read in all 3 recording sessions for a total of 6 repetitions of each sentence and 30 sentences overall. The sentences are "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." The read sentences were used to test phonetic content variability.

The pet-directed speech was recorded to represent an extreme speaking style. The speakers were instructed to speak affectionately for at least 1 minute to small pets displayed in a video. Resulting utterances were similar to infant-directed speech, which is characterized by exaggerated prosody [16]. Within the database, the read sentences and pet-directed speech represented contrasting speaking styles and thus were suitable for examining the effect of varying speaking style. In this study, all utterances were downsampled to 8 kHz for consistency with general telephone-channel speaker recognition tasks.

2.2. Feature Selection

Previously, we showed that features inspired by psychoacoustic representations of voice quality [17, 18] are effective for automatic speaker verification and human response modeling [14]. This feature set included F0, F1, F2, F3, H1*-H2* (the difference between the first and second harmonic amplitudes), H2*-H4* (the difference between the second and forth harmonic amplitudes), H4*-H2k* (the difference between the fourth harmonic amplitude and the amplitude of the harmonic component near 2 kHz), and cepstral peak prominence (CPP, [19]). The asterisks (*) indicate that the formant effect on a harmonic amplitude was corrected [20]. The amplitude difference between the harmonic component near 2 kHz and 5 kHz (H2k*-H5k) was not included because the speech samples were bandlimited to 4 kHz. This set of parameters will be denoted as the *VQual1* features throughout this paper.

Two considerations were made to improve the voice quality feature set. One consideration was whether formant correction would be performed when estimating harmonic amplitude differences. It has been observed that over-correction may occur when formants are situated on top of a harmonic. Such inaccuracies may weaken the ability of the features to separate speakers. Also, uncorrected harmonic amplitudes might be another way of representing formant differences between speakers.

The second consideration was which form of the formant amplitudes would be added. Formant amplitudes are frequently used to represent voice quality, and they may have important speaker-specific information. Frequently used features include H1*-A1*, H1*-A2*, and H1*-A3* [21, 22] where A1, A2, and A3 are the amplitudes of the first, second, and third formants. All features mentioned above, as well as A1, A2, and A3, were chosen as candidate features.

The features were extracted every 10 msec using VoiceSauce software [23], with Praat [24] chosen as the method for extracting pitch and formant frequencies. The ability of each candidate feature to separate speakers was examined using the F-ratio [25, 26] separability measure, defined as:

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

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.

Four conditions were examined using the UCLA database. In the first ("mixed phonetic content"), all read sentences were randomly distributed to 5 subsets, and the average F-ratio was found. In the second ("separated phonetic content"), the F-ratios were calculated within 5 subsets of the same sentences and averaged. Two analogous conditions ("mixed style" and "separated style") were also created. The difference between the content conditions and the difference between the style conditions are shown in Figure 1. Two different "separated style" conditions are displayed using just read sentences and just pet-directed speech.

The features with high F-ratios were similar between the content and style variability cases. This is partially due to the fact that content variability is also present in the style variability subset. The harmonic amplitude difference features had higher F-ratios without formant correction than with correction in general. Thus, the new feature set replaced H1*-H2*, H2*-H4*, and H4*-H2k* with H1-H2, H2-H4, and H4-H2k. Also, as the raw formant amplitudes generally had the highest F-ratios out of all the features using formant amplitudes, A1, A2, and A3 were added to the feature set. The final chosen feature set included F0, F1, F2, F3, H1-H2, H2-H4, H4-H2k, CPP, A1, A2, and A3. These features will be called the *VQual2* features and are used throughout the ASV experiments. We expected that the VQual2 features would better separate speakers than the VQual1 features and would further improve ASV system performance.

Interestingly, the F-ratios in the pet-directed subset were high among the style variability conditions. This is likely because each speaker had his or her own unique style of talking to pets, making speakers more distinct and increasing between-speaker variability. This uniqueness in speaking style between speakers will likely result in decent performance in same-style ASV experiments which will be discussed in Section 3.

Since the F-ratio analyzes individual features, the ability of the entire MFCC and VQual2 feature set was analyzed using the J-measure [27, pp. 280–283], defined as:

$$J = \mathrm{Tr}\{S_w^{-1} S_m\} \quad (2)$$

where $S_m = S_w + S_b$ is the mixture scatter matrix and S_w, S_b are the within-class scatter matrix and the between-class scatter

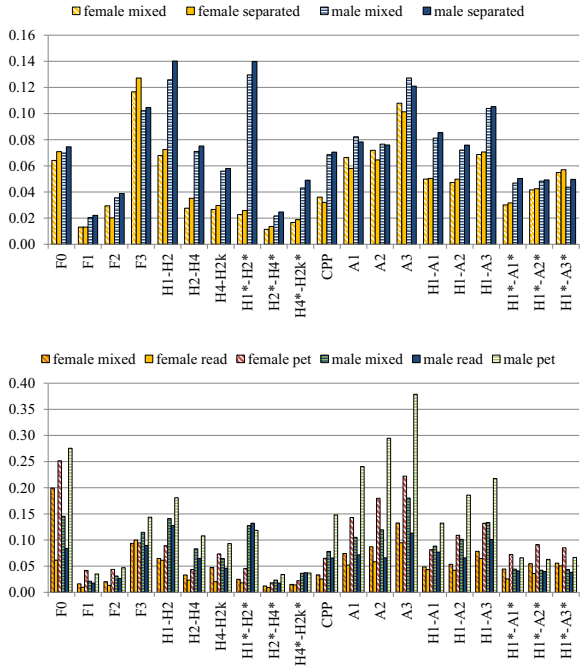

Figure 1: *Computed F-ratios of various voice quality features using the UCLA Speaker Variability Database for content (top) and style (bottom) variability. "Separated" in the top panel indicates the subsets were separated by sentences. "Read" and "pet" in the bottom panel indicate the subsets only contained read sentences or pet-directed speech.*

matrix, respectively. The J-measures using the read sentences in the UCLA database are shown in Figure 2. As the voice quality features are added to MFCCs, the J-measure increases. Thus, we hypothesize that the VQual2 features provide complementary information to the MFCCs. This will be evaluated in the next couple of sections.

3. ASV Speaker Variability Analysis

A state-of-the-art i-vector[28]/PLDA[29] ASV system was used for the following experiments. This system was trained on the NIST SRE04, 05, 06, and 08 databases. MFCCs of dimension 20, along with first derivatives, were used as baseline features. The second derivatives were not used because they did not provide significant performance gain. The VQual1 and VQual2 features, along with first and second derivatives, were used as alternative feature sets. After obtaining the PLDA scores from each system, score fusion was used for further improvements. Since the VQual2 features performed better than the VQual1 features in almost all cases, only fusion performance with MFCCs and the VQual2 features is reported. The system outputs were linearly combined using

$$s = \alpha s_v + (1 - \alpha) s_m \tag{3}$$

where s_m is the PLDA score using MFCCs, s_v is the PLDA score using VQual2 features, and α is the coefficient of s_v chosen from the range of 0 to 1. PLDA scores using both MFCCs and VQual2 features were first scaled to have zero-mean and unit-variance before fusion was performed.

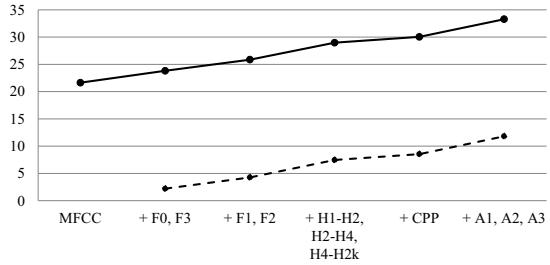

Figure 2: *Computed J-measures of VQual2 and MFCC features using read sentences (solid line). Each point on the chart is the J-measure of all features below and to the left of that point. The dashed line shows VQual2 features without MFCCs for comparison.*

3.1. Content Variability

The effect of varying phonetic content was analyzed with the five sentences from 100 female and 100 male speakers from the UCLA database. Two different conditions were compared: same-text trials and different-text trials. In both conditions, one sentence was chosen to enroll speakers, and one sentence was chosen to test the system. In the "same-text" trials, the same sentence was used for enrollment and testing. In the "different-text" trials, the sentences used for enrollment and testing were different. All possible combinations of enrollment–test utterance pairs were used except for the case when the enrollment and test utterances were identical. The overall performance in terms of EER is shown in Table 1.

As predicted, system performance degraded severely when content variability was large. Using MFCCs, error rates increased dramatically for both females and males comparing same and different-text trials. The VQual2 features did not perform as well as MFCCs in the same-text trials. In the different-text trials, the VQual2 features performed almost as well as MFCCs for females and even better than MFCCs for males. In all cases, fusion with the VQual2 scores provided relative improvements of at least 13.97%, suggesting that the VQual2 features contain complementary information to MFCCs.

For comparison, Table 1 also includes results from human perception experiments presented in a previous study [14], where the listeners were asked to determine if a given pair of read sentences was spoken by the same speaker or two different speakers. Note that the human listeners were not affected by the content difference as much as the ASV system.

3.2. Style Variability

The effect of varying speaking style was examined with both read sentences and pet-directed utterances. The read sentences were randomly concatenated together for each speaker until 5 seconds of speech were collected. This was done twice to create enrollment and testing sets for every speaker. No utterance was used for both enrollment and testing. The pet-directed speech was cut into two non-overlapping segments containing 5 seconds of speech for enrollment and testing. One female and six male speakers were removed due to poor quality or low amounts of speech in the pet-directed recordings.

Four types of trials were designed, denoted as "enrollment data–testing data": read–read, read–pet, pet–read, and pet–pet. In the read–read and pet–pet trials, the same speaking style was

Table 1: *ASV system performance in terms of EER (%) with content and style variability using the UCLA database. The fusion performance is obtained by fusing the PLDA scores from the MFCCs and VQual2 features, and its relative improvement compared to MFCCs is denoted in parentheses. The human speaker recognition result from a previous study [14] is added in the last row for comparison.*

	female		male		female			male		
	same -text	different -text	same -text	different -text	read –read	pet –pet	read –pet	read –read	pet –pet	read –pet
MFCC	7.71	28.14	5.97	28.33	3.65	19.19	30.30	2.13	6.38	19.30
VQual1	15.66	31.48	15.15	28.33	10.10	18.18	36.77	5.48	12.72	30.85
VQual2	12.67	28.23	13.63	27.73	6.40	18.18	35.06	3.19	11.70	28.72
Fusion	6.22	24.21	4.93	23.07	3.03	12.79	29.29	1.06	4.25	19.15
(% imp.)	(19.33)	(13.97)	(17.42)	(16.80)	(16.99)	(33.35)	(3.33)	(50.23)	(33.39)	(0.78)
Human [14]	10.00	12.22	-	-	-	-	-	-	-	-

chosen for both enrollment and testing. In the read–pet trial, read sentences were used for enrollment and pet-directed speech was used for testing. The pet–read trial was opposite to this. The results are shown in Table 1. The results from pet–read trials are omitted as they were almost identical to read–pet trials.

As expected, the error using MFCCs increased dramatically both for female and male speakers in the style-mismatched condition compared to the read–read trials. This might be due not only to feature distortion by the exaggerated prosody but also to limited phonetic content in the pet-directed speech samples. Most speakers had very limited phonetic content, often speaking incomprehensibly or repeating phrases such as "Aww!" and "So cute!". Score fusion with the VQual2 features showed decent improvements for the style-matched conditions, notably 33% improvement for the pet-pet trials. However, there was little improvement for the style-mismatched condition. This is expected since the VQual2 features themselves partially reflect the speaking style of a speaker and are susceptible to style changes.

4. ASV Short-Utterance Evaluation

Using the same ASV system described in Section 3, a series of experiments was conducted using the NIST SRE10 database condition 5 extended task [30] for evaluation. In addition to the full utterance (about 5 min) evaluation, new enrollment and testing datasets were created using the SRE10 data by cutting the utterances to contain 10, 5, and 2 seconds of speech. The performance in terms of EER and the optimal choice of α are shown in Table 2, where α is the weight of the VQual2 scores in the linear score fusion as in Eq. (3).

The VQual2 features showed around 2% absolute improvement compared to the VQual1 features in all conditions. As expected, both MFCCs and VQual2 features performed worse as utterances became shorter. However, the VQual2 features were able to improve the performance of the system through score fusion by providing complementary information to MFCCs. The weight of the VQual2 scores in the fusion (α) increased as utterances became shorter, suggesting that the system relied more on the VQual2 scores as the amount of speech decreased.

Table 2: *ASV system performance in terms of EER (%) with the NIST SRE10 database. The relative improvement by fusion (MFCC+VQual2) is noted in parentheses and optimal coefficient α is shown.*

	full	10-10	5-5	2-2
MFCC	2.89	10.88	16.90	28.47
VQual1	8.96	19.60	25.18	32.95
VQual2	7.91	17.23	22.82	30.92
Fusion	2.80	9.53	14.91	25.95
(% imp.)	(3.11)	(12.41)	(11.78)	(8.85)
α	.10	.29	.35	.46

5. Conclusion

This study analyzed ASV system performance when content and style variability were large, especially when utterances were short. Initial examination with the UCLA Speaker Variability Database was conducted. A VQual2 feature set was chosen that effectively separated speakers in these large variability conditions using the F-ratio and J-measure.

ASV experiments using an i-vector/PLDA ASV system with MFCCs and VQual2 features were conducted, along with linear score fusion. It was found that when content and/or style conditions were mismatched, the system error rate increased significantly. In the mismatched content conditions, score fusion with VQual2 features improved performance. But unlike humans, there was still a large difference in performance between matched and mismatched content conditions, so there is much to be done in future work. In the mismatched style conditions, score fusion did not improve performance much. However, the performance gain in the pet–pet trials suggests that voice quality features might be able to capture speakers' idiosyncratic way of exaggerating prosody.

To examine the effects of using voice quality features with a larger set of utterances and speakers, ASV experiments were conducted using the NIST SRE10 telephone speech database. Results suggest that short utterances ($<$ 10 sec) benefit from using VQual2 features more than long utterances (about 5 min).

Future studies will examine additional features such as prosodic and subglottal features. We will also investigate how phonetic content, speaking style, and affect can influence speaker verification tasks by human listeners, and contrast the different acoustic cues and recognition strategies used by humans and machines.

6. Acknowledgements

We thank John Hopkins University Human Language Technology Center of Excellence (JHU HLTCOE) for providing the i-vector/PLDA system and computational resources. This research was supported in part by the NSF.

7. References

[1] R. K. Das, S. Jelil, and S. R. M. Prasanna, "Significance of Constraining Text in Limited Data Text-Independent Speaker Verification," in *Proc. International Conference on Signal Processing and Communications (SPCOM)*, 2016, pp. 1–5.

[2] W. B. Kheder, D. Matrouf, M. Ajili, and J.-F. Bonastre, "Probabilistic Approach Using Joint Long and Short Session i-vectors Modeling to Deal with Short Utterances for Speaker Recognition," in *Proc. Interspeech*, San Francisco, USA, 2016, pp. 1830–1834.

[3] 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, 2015.

[4] C. Y. Espy-Wilson, S. Manocha, and S. Vishnubhotla, "A New Set of Features for Text-Independent Speaker Identification," in *Proc. Interspeech*, Pittsburgh, USA, 2006, pp. 1475–1478.

[5] J. Gudnason and M. Brookes, "Voice Source Cepstrum Coefficients for Speaker Identification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4821–4824.

[6] 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, pp. 1–19, 2015.

[7] R. K. Das and S. R. Mahadeva Prasanna, "Exploring Different Attributes of Source Information for Speaker Verification with Limited Test Data," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 184–190, 2016.

[8] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980.

[9] S. R. Schweinberger, H. Kawahara, A. P. Simpson, V. G. Skuk, and R. Zäske, "Speaker Perception," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 1, pp. 15–25, 2014.

[10] E. S. Segundo, P. Foulkes, and V. Hughes, "Holistic Perception of Voice Quality Matters more than L1 when Judging Speaker Similarity in Short Stimuli," in *Proc. Australasian Conference on Speech Science and Technology (SST)*, Parramatta, Australia, 2016, pp. 309–312.

[11] V. Hautamäki, T. Kinnunen, M. Nosratighods, K.-A. Lee, B. Ma, and H. Li, "Approaching Human Listener Accuracy with Modern Speaker Verification," in *Proc. Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 1473–1476.

[12] R. Roebuck and J. Wilding, "Effects of Vowel Variety and Sample Length on Identification of a Speaker in a Line-up," *Applied Cognitive Psychology*, vol. 7, no. 6, pp. 475–481, 1993.

[13] J. Kreiman, S. J. Park, P. A. Keating, and A. Alwan, "The Relationship Between Acoustic and Perceived Intraspeaker Variability in Voice Quality," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 2357–2360.

[14] S. J. Park, C. Sigouin, J. Kreiman, P. Keating, J. Guo, G. Yeung, F.-Y. Kuo, and A. Alwan, "Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition," in *Proc. Interspeech*, San Francisco, USA, 2016, pp. 1044–1048.

[15] IEEE Subcommittee on Subjective Measurements, "IEEE Recommended Practices for Speech Quality Measurements." *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 297, pp. 227–246, 1969.

[16] D. Burnham, C. Kitamura, and U. Vollmer-Conna, "What's New, Pussycat? On Talking to Babies and Animals," *Science*, vol. 296, no. 5572, p. 1435, 2002.

[17] J. Kreiman and B. R. Gerratt, "Perceptual Interaction of the Harmonic Source and Noise in Voice," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 492–500, 2012.

[18] M. Garellek, R. Samlan, B. R. Gerratt, and J. Kreiman, "Modeling the Voice Source in Terms of Spectral Slopes," *The Journal of the Acoustical Society of America*, vol. 139, no. 3, pp. 1404–1410, 2016.

[19] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic Correlates of Breathy Vocal Quality," *Journal of Speech Language and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[20] M. Iseli, Y.-L. Shue, and A. Alwan, "Age, Sex, and Vowel Dependencies of Acoustic Measures Related to the Voice Source," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2283–2295, 2007.

[21] H. M. Hanson, "Glottal Characteristics of Female Speakers: Acoustic Correlates," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466–481, 1997.

[22] J. Vaňková and R. Skarnitzl, "Within- and Between-Speaker Variability of Parameters Expressing Short-Term Voice Quality," in *Proc. Speech Prosody*, 2014, pp. 1081–1085.

[23] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for Voice Analysis," in *Proc. International Congress of Phonetic Sciences (ICPhs) XVII*, Hong Kong, 2011, pp. 1846–1849.

[24] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer," 2017. [Online]. Available: http://www.praat.org/

[25] S. Nicholson, B. Milner, and S. Cox, "Evaluating Feature Set performance using the F-ratio and J-measures," in *Proc. Eurospeech*, 1997, pp. 413–416.

[26] X. Lu and J. Dang, "An Investigation of Dependencies Between Frequency Components and Speaker Characteristics for Text-Independent Speaker Identification," *Speech Communication*, vol. 50, no. 4, pp. 312–322, 2008.

[27] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Elsevier, 2009.

[28] 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, 2011.

[29] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for Speaker Verification with Utterances of Arbitrary Duration," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7649–7653.

[30] National Institute of Standards and Technology, "The NIST Year 2010 Speaker Recognition Evaluation Plan," 2010.