Interspeech 2018
2-6 September 2018, Hyderabad

Using Voice Quality Supervectors for Affect Identification

Soo Jin Park, Amber Afshan, Zhi Ming Chua, and Abeer Alwan

Electrical and Computer Engineering Department, University of California Los Angeles, USA

sj.park@ucla.edu, amberafshan@g.ucla.edu

Abstract

The voice quality of speech sounds often conveys perceivable information about the speaker's affect. This study proposes perceptually important voice quality features to recognize affect represented in speech excerpts from individuals with mental, neurological, and/or physical disabilities. The voice quality feature set consists of F0, harmonic amplitude differences between the first, second, fourth harmonics and the harmonic near 2 kHz, the center frequency and amplitudes of the first 3 formants, and cepstral peak prominence. The feature distribution of each utterance was represented with a supervector, and the Gaussian mixture model and support vector machine classifiers were used for affect classification. Similar classification systems using the MFCCs and ComParE16 feature set were implemented. The systems were fused by taking the confidence mean of the classifiers. Applying the fused system to the Interspeech 2018 Atypical Affect subchallenge task resulted in unweighted average recalls of 43.9% and 41.0% on the development and test dataset, respectively. Additionally, we investigated clusters obtained by unsupervised learning to address gender-related differences.

Index Terms: emotion recognition, human-computer interaction, computational paralinguistics

1. Introduction

Speech emotion recognition (SER) is a fast growing research area having a broad range of potential applications. For example, SER can be used for developing applications which can detect and react to a user's emotion. Such an application might be an effective tool to support individuals with disabilities or with mental disorders by automatically diagnosing or monitoring their mental states. In addition, if a system can automatically assess how speakers' affect are perceived by others, it can, through feedback, help patients interact with other people effectively. This study aims to automatically classify perceived affect represented in speech samples from individuals with disabilities, as a participation in the Interspeech 2018 Atypical Affect subchallenge [1]. The dataset used in this challenge consists of spontaneous speech samples from 15 speakers with mental, neurological, and/or physical disorders. The affect labels for the samples were made by human listeners through a gamified crowdsourcing platform [2].

The performance of SER systems depends highly on which acoustic features are used [3, 4]. In this study, we propose to use a set of features which was inspired by a psychoacoustic model comprised of acoustic features that account for perceived voice quality [5]. Voice quality has been frequently associated with affect [6, 7], and a strong relationship between voice quality and perceived affect was found through experimental studies with human listeners [8]. Yet, it has been difficult to automatically and reliably extract acoustic features related to voice quality. One approach is based on the assumption that voice quality can be better measured by estimating the glottal source signal, and often involves inverse-filtering. However, the reliability of automatic source estimation is limited [9]. Other techniques have been developed to estimate the parameters that represent voice quality directly from the signal (e.g. jitter and shimmer [10]). Still, the relationship between such parameters and perceptual responses is unclear [11]. The proposed feature set in this study, on the other hand, is based on a perceptually valid model. Thus, we expect the feature set to be correlated with perceived affect, and hence may improve automatic classification of perceived affect.

The supervector framework [12] is utilized in this study to model the distribution of acoustic features within an utterance. In this framework, each utterance is represented with a single vector that is constructed by concatenating the mean vectors of a Gaussian mixture model representing the feature distribution within an utterance. The mixture model is often adapted from the universal background model (UBM), which is a statistical model for speech sounds, usually trained with a large amount of recordings from a large number of speakers. A technique derived from this approach, the i-vector, effectively represents an utterance in a low-dimensional subspace [13]. The i-vector framework is most effective when a large database including a wide range of affect and speaker variability is available. Considering that the available amount of data was limited to train both the UBM and the i-vector subspaces, we decided to directly use supervectors for this task.

Compensating for the effect of a particular disorder on speech signals was not explicitly attempted in this study because the available metadata did not include information to perform such analysis. Instead, we focused on predicting perceived affect regardless of the kind of disorder.

The rest of the paper is organized as follows. The feature sets and the utterance representation method used in this study are presented in Section 2 and in Section 3, respectively. The classification and fusion techniques to build the complete system are described in Section 4, and system performance is evaluated in Section 5. The paper concludes in Section 6.

2. Acoustic Features

2.1. VQual: Voice Quality Features

The feature set proposed in this study is inspired by a psychoacoustic model of voice quality [5, 14]. The model describes voice quality with acoustic parameters including the fundamental frequency (F0), harmonic-to-noise ratio, and differences in harmonic amplitudes. Extensive studies (e.g., [15, 16, 17]) have shown that listeners are perceptually sensitive to all these parameters, and that, as a set, the parameters are sufficient to quantify source contributions to voice quality. Thus, the model can be considered perceptually valid. These acoustic parameters can be estimated directly from the speech signal without inverse filtering, and an automatic estimation algorithm is avail-

10.21437/Interspeech.2018-1401

able from the VoiceSauce toolkit [18].

These voice quality parameters were utilized to represent speaker identity, and improved automatic speaker verification system performance [19, 20, 21]. The feature set used in this study, denoted as VQual, included F0, F1, F2, F3, harmonic amplitude differences H1-H2, H2-H4, H4-H2k, formant amplitudes A1, A2, A3, and cepstral peak prominence (CPP, [22]). Here, H1, H2, H4, and H2k indicate the amplitudes of first, second, fourth harmonics, and the harmonic nearest to 2 kHz. The first and second derivatives of these features were also used.

2.2. Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCCs) represent the overall spectral envelope of the speech signal. They are closely related to the phonetic information in speech at the frame level, but physiological changes related to affect might shape the vocal tract in a certain way that can be reflected in the spectral envelope. Thus, the MFCC distribution over an utterance might have information of the emotional state of the speaker. MFCCs of dimension 20, including the zeroth coefficient, were extracted using a 25 msec window and a 10 msec frame shift. The first and second derivatives were also computed.

2.3. ComParE16 Feature Set

The computational paralinguistics challenge provided a baseline feature set [23, 24] that can be extracted using the OpenSMILE toolkit [25]. The set consists of F0, energy, spectral, cepstral coefficients and voicing-related frame-level features which are referred to as low-level descriptors. They also include the zero-crossing rate, jitter, shimmer, harmonic-to-noise ratio, spectral harmonicity and psychoacoustic spectral sharpness. The complete feature set had a dimension of 65.

3. Utterance Representation

3.1. Supervector Construction

A universal background model (UBM) was constructed using the 'neutral' class of the AtypicalAffect dataset, and the data provided for the SelfAssessedAffect subchallenge. The two datasets were in the German language and contained spontaneous speech. After the UBM was trained, the feature distribution of each utterance was modeled with a Gaussian mixture model (GMM) adapted from the UBM using the maximum a posteriori criterion [12]. The supervector, which is the concatenated mean vectors of the GMM, was used to represent each utterance.

3.2. Utterance Clustering based on F0

F0 distributions for the training and development datasets were bimodal, as shown in the first column of Figure 1. The bimodal distribution suggests the possible effect of gender differences (females having, on average, a higher F0 than males). It was also observed that the F0 distribution for the training and development datasets were not similar. For example, there were two peaks in the F0 distribution of the 'happy' class in the training dataset, while the F0 distribution of the the same class in the development dataset did not show a clear second peak. The mismatch in the F0 distribution between the datasets suggests that the gender distribution in the datasets might differ significantly. Considering that affect representation can be different across gender [26], gender distribution mismatch could degrade classification performance. For example, if the majority of the

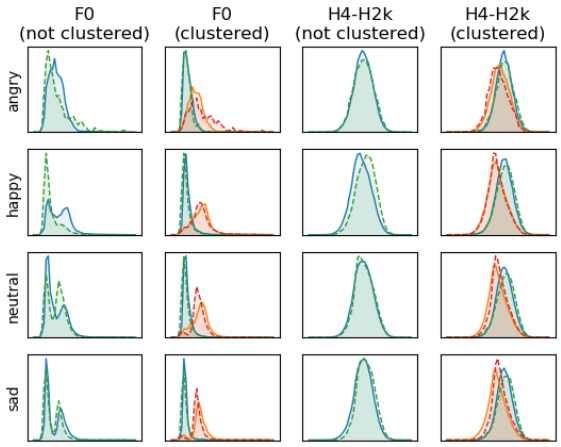


Figure 1: *Example feature distributions without and with unsupervised clustering. The distribution for the training dataset (solid line) and the development dataset (dashed line) are shown separately. The distributions within clustered subsets match better between the training and development datasets compared to the non-clustered distributions.*

training data were from males while the development data had more females than males, the statistical model built on the features extracted from the training data might not be able to accurately predict the affect for the development data.

One possible approach to alleviate the mismatch problem is gender-dependent modeling, However, gender labels were not available, hence, we used F0 to group the utterances into two clusters. A Gaussian mixture model with two mixtures was used for clustering, based on the median value of F0 within each utterance. The resulting cluster size differed significantly across datasets. For example, in the training dataset, the number of utterances for the 'happy' class was 165 and 578 in low-F0 cluster and high-F0 cluster, respectively, while in the development dataset, the corresponding numbers were 632 and 333.

Example feature distributions (F0 and H4-H2k) without and with the clustering are shown in Figure 1. Note that clustering was performed at the utterance level, while the feature distribution is computed at the frame level. Because utterances can have low-F0 frames while the utterance F0 median value is high, frame-level F0 distribution might not show a clear separation between clusters. Clustering resulted in a more matched distribution between the training and development datasets, especially for 'happy' and 'angry' F0 distributions, and for the H4-H2k distribution in the 'happy' class.

However, this clustering inevitably results in a reduced amount of data for training within each cluster, which might introduce limitations to classification performance.

4. Affect Classification

4.1. Classifiers

4.1.1. The Gaussian Mixture Model Classifier

Because the number of utterances in each affect class is limited, training the Gaussian mixture models (GMMs) directly from the samples within each class is prone to overfitting. In order to mitigate the overfitting problem, a GMM was trained using the 'neutral' class, not only because it is the class with the highest

number of samples, but also because emotional speech could be regarded as a variation of neutral speech. The models for the remaining three classes ('sad', 'angry', and 'happy') were adapted from the 'neutral' model. The classification decision was made based on the log-likelihood that a test supervector was drawn from each class.

4.1.2. The Support Vector Machine Classifier

A support vector machine (SVM) using a linear kernel implemented in the Weka toolkit [27] was used. The supervector configuration that performed the best for each feature set was used for the SVM classification. The complexity parameter C was chosen between the values $10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$ and 10^{-6}, so that it maximizes the system performance on the development dataset for each feature set. Data upsampling was carried out for the under-represented classes to address the data imbalance problem.

4.2. System Fusion

The best performing configuration for each feature set/classifier combination on the development dataset was selected as the representative system for that combination. The system fusion was performed on the n-best performing representative systems.

In order to fuse the results from the GMM and SVM classifiers, the log-likelihood output from the GMM classifier was converted into the confidence score so that it was consistent with the baseline SVM classifier results. The SVM classifier's confidence score was the probability that the test utterance belonged to a class, and it was calculated so that scores for the classes added up to one. On the other hand, the GMM likelihood was calculated for each class independently, hence, there was no guarantee that the likelihoods for the classes added up to one. Thus, the confidence for the i-th class, c_i, was calculated as follows:

$$c_i = \frac{\exp(l_i - \mu)}{\sum_{i=1}^{M} \exp(l_i - \mu)} \quad (1)$$

where l_i is the log-likelihood for the i-th class, M is the number of classes, and $\mu = 1/M \sum_{i=1}^{M} l_i$ is the mean of the log-likelihoods across the classes. The confidence scores from the classifiers were averaged and used for the combined class decision.

5. Experimental Results

5.1. Individual Supervector-Based System Performance

The performances of individual systems in terms of unweighted average recall (UAR) are summarized in Table 1. VQual performed better than both MFCCs and the baseline ComParE16 feature set in all conditions.

Contrary to expectation, the clustering did not provide a performance gain. For the MFCC feature set, the performance degraded from 41.37% UAR to 36.78%. The degradation might be due to overfitting because of insufficient amount of data to train within each cluster. Because MFCCs had 60-dim features while VQual had 33-dim features, the shortage of data points might have affected the MFCC-based system more critically.

It is interesting to note that the ComParE16 feature set performance was improved to 40.7% UAR by using the supervector, compared to the OpenSMILE baseline system with a UAR of 37.8%; recall that the baseline system uses a statistics vector for utterance representation. Because both systems used the same acoustic feature set, these results can be used to compare the effect of different methods in modeling the utterances.

Table 1: *Individual system performance in terms of unweighted average recall (UAR, [%]). The performance was measured on the development dataset. The system configurations chosen for system fusion are denoted with asterisks (*), and the ranking among them is shown in the last column. The SVM parameter $C = 10^{-6}$, 10^{-5}, and 10^{-3} was used for the VQual, MFCCs and ComParE16 features, respectively.*

feature set	clustering	classifier	UAR	ranking
VQual	Yes	GMM	39.40	-
VQual	No	GMM	*41.37	2
VQual	No	SVM	*41.92	1
MFCC	Yes	GMM	36.78	-
MFCC	No	GMM	*41.21	3
MFCC	No	SVM	*40.95	4
ComParE16	Yes	GMM	36.19	-
ComParE16	No	GMM	*40.21	6
ComParE16	No	SVM	*40.71	5

Table 2: *Fused system performance on the development dataset, in terms of unweighted average recall (UAR, [%]). The best performing combination is boldfaced.*

		+ baseline
2-best	41.71	42.24
3-best	42.60	**43.92**
4-best	43.89	43.78
5-best	**44.42**	42.96
6-best	42.69	41.04

5.2. Fused System Performance

The configurations selected for each feature set/classifier combination are denoted with asterisks (*), and their performance ranking is shown in Table 1. The n-best system fusion performance is shown in the first column of Table 2. The two best systems were both VQual-based systems, one with an SVM and the other with a GMM classifier. The fact that both systems used the same acoustic information might be the reason why fusion did not improve performance. Adding the third and the forth best system, which were based on MFCCs, the UAR improved by 2.18%, providing complementary information to VQual-based systems. The 5-best system combination, by adding the ComParE16/SVM system, performed the best (UAR=44.42%).

The OpenSMILE baseline system, with a UAR of 37.8%, used different utterance representation from the supervector framework. Even though the performance was lower than the systems introduced in this study, the baseline system might be complementary. Thus, the fusion of the baseline system in addition to the n-best systems was investigated. The performance with the baseline system is shown in the second column of Table 2. Fusing the baseline system with the 2 and 3 best systems improved the performance, suggesting a complementary effect. However, fusing it with the 4, 5 and 6 best systems degraded the system performance.

5.3. System Performance Evaluation on the Test Dataset

The complete system block diagram is shown in Figure 2 and its performance on the development and test dataset is reported

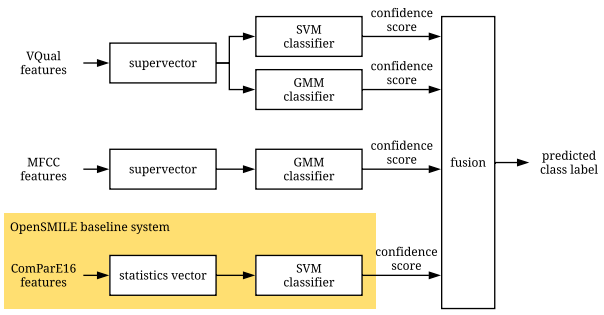
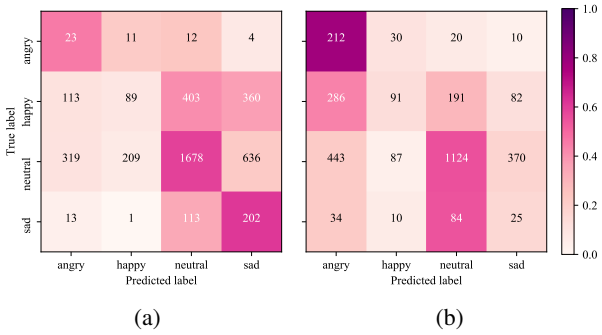


Figure 2: *The complete system block diagram.*



(a) (b)

Figure 3: *Confusion matrices for the results from (a) the development dataset and (b) the test dataset.*

in Table 3. As the number of evaluation trials reached the limit, the best performing system on the development dataset could not be evaluated on the test dataset. However, the evaluation result on the test dataset is available for the second best system, which was the fusion of OpenSMILE baseline and the 3 best systems.

Table 3: *System performance in terms of unweighted average recall (UAR, [%]) on the development and test datasets.*

	UAR devel	UAR test
OpenSMILE baseline	37.8	43.1
OpenSMILE + 3-best	43.9	41.0

The proposed system significantly outperformed OpenSMILE baseline system on the development dataset: the system performance improved from 37.8% to 43.9%. On the test dataset, the proposed system did not show similar trends: its UAR was 41.0%, while the baseline was at 43.1%.

Confusion matrices of the proposed system for the development and test datasets are shown in Figure 3. For the development dataset, the recall improvement was evident for the 'angry' and 'sad' classes compared to the OpenSMILE baseline. The 'angry' and 'sad' recalls improved from 30.00% to 46.00%, and from 43.16% to 61.40%, respectively. The precisions for those classes showed small difference between the baseline and the proposed systems compared to the recall improvements: the 'angry' class precision slightly increased from 4.35% to 4.91%, whereas the 'sad' class precision decreased from 17.77% to 16.81%.

The performance pattern, unfortunately, was not consistent in the test dataset. For the test dataset, the 'angry' recall and precision increased to 77.94% and 21.74%, respectively. However, the 'sad' recall and precision decreased to 16.34% and 5.13%, respectively. Because those two had the least amount of data in the training dataset, overfitting might have yielded these results. For example, there were only 125 'angry' voices and 187 'sad' voices while there were 2,287 'neutral' voices in the training dataset. Thus, it is likely that the two classes did not have sufficient data to construct reliable models.

For both datasets, the 'happy' class was the least recalled class. One possible explanation for this confusion is the mismatch across the training and development datasets as observed in Section 3.2.

6. Conclusion

A perceptually valid voice quality feature set (VQual) was applied to affect classification on the recordings collected from individuals with mental, neurological, and/or physical disorders. Compared to the baseline ComParE16 feature set, VQual showed better affect classification performance in all experimental configurations. The VQual feature set also performed equivalently well, or better in some configurations, compared to MFCCs. It is noteworthy that the proposed VQual feature set had lower dimension than the MFCCs and the baseline ComParE16 feature set, but it could outperform those feature sets.

The supervector approach used in this study showed its effectiveness in representing the utterances. The utterance-level distribution of VQual features and MFCCs was effectively modeled with this approach, resulting in the system outperforming the OpenSMILE baseline system on the development dataset. Additionally, using a supervector derived from ComParE16 feature set resulted in a better performance than the baseline system which used the statistics vector for the same feature set. These results suggest that in the cases when the amount of data is insufficient to apply the i-vector framework, the supervector approach can be a viable alternative to represent the local feature distribution within an utterance.

The confidence score that an utterance was drawn from a class was used for system fusion. When the systems using different features were fused, the performance improved, suggesting complementary effect between feature sets. The system fusion configuration was finalized based on the single system performance, and the complete system performance was analyzed based on confusion matrices. The performance gain was obtained by improving the recall for 'sad' and 'angry' classes. However, the proposed system was less effective on the test dataset, suggesting overfitting due to insufficient amounts of data especially in the 'sad' and 'angry' classes.

Analysis of the system performance suggests that further improvements could be made by better modeling the classes with limited training data. Addressing acoustic mismatch across datasets would be another important direction for future studies.

7. References

[1] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-m. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Chal-

lenge: Atypical & Self-Assessed Affect, Crying & Heart Beats," in *Proc. Interspeech*, Hyderabad, India, 2018.

[2] S. Hantke, H. Sagha, N. Cummins, and B. Schuller, "Emotional Speech of Mentally and Physically Disabled Individuals: Introducing the EmotAsS Database and First Findings," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3137–3141.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[4] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and Classifiers for Emotion Recognition from Speech: A Survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.

[5] J. Kreiman, B. R. Gerratt, M. Garellek, R. Samlan, and Z. Zhang, "Toward a Unified Theory of Voice Production and Perception," *Loquens*, vol. 1, no. 1, pp. 1–9, jun 2014.

[6] K. R. Scherer, "Vocal Affect Expression: A review and a Model for Future Research," *Psychological Bulletin*, vol. 99, no. 2, pp. 143–165, 1986.

[7] I. R. Murray and J. L. Arnott, "Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[8] C. Gobl and A. Ní Chasaide, "The Role of Voice Quality in Communicating Emotion, Mood and Attitude," *Speech Communication*, vol. 40, pp. 189–212, 2003.

[9] M. Fröhlich, D. Michaelis, and H. W. Strube, "SIM-Simultaneous Inverse Filtering and Matching of a Glottal Flow Model for Acoustic Speech Signals," *The Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 479–488, 2001.

[10] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman, "Stress and Emotion Classification using Jitter and Shimmer Features," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, Hawaii, USA, 2007, pp. IV–1081–IV–1084.

[11] J. Kreiman and B. R. Gerratt, "Jitter, shimmer, and noise in pathological voice quality perception," in *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03), ISCA Tutorial and Research Workshop*, Geneva, Switzerland, 2003, pp. 57–62.

[12] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[13] 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.

[14] 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.

[15] J. Kreiman and B. R. Gerratt, "Perceptual Sensitivity to First Harmonic Amplitude in the Voice Source," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2085–2089, 2010.

[16] ——, "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.

[17] M. Garellek, R. A. Samlan, J. Kreiman, and B. R. Gerratt, "Perceptual Sensitivity to a Model of the Source Spectrum," in *Proc. Meetings on Acoustics*, vol. 19, Montreal, Canada, 2013, pp. 060 157–060 157.

[18] Y.-L. Shue, P. A. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A Program for Voice Analysis," in *Proc. International Congress of Phonetic Sciences (ICPhs) XVII*, vol. 126, Hong Kong, 2011, pp. 1846–1849.

[19] 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.

[20] S. J. Park, C. Sigouin, J. Kreiman, P. A. 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, sep 2016, pp. 1044–1048.

[21] S. J. Park, G. Yeung, J. Kreiman, P. A. Keating, and A. Alwan, "Using Voice Quality Features to Improve Short-Utterance, Text-Independent Speaker Verification Systems," in *Proc. Interspeech*, Stockholm, Sweden, aug 2017, pp. 1522–1526.

[22] 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.

[23] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proc. Interspeech*, 2016, pp. 2001–2005.

[24] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common," *Frontiers in Psychology*, vol. 4, pp. 1–12, 2013.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.

[26] T. Vogt and E. André, "Improving Automatic Emotion Recognition from Speech via Gender Differentiation," in *Proc. Language Resources and Evaluation Conference*, Genoa, Italy, 2006, pp. 1123–1126.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.