Zewoudie, Abraham; Bäckström, Tom

# Voice-Quality Features for Replay Attack Detection

*Published in:*
2022 30th European Signal Processing Conference (EUSIPCO)

Published: 01/01/2022

*Document Version*
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

*Please cite the original version:*
Zewoudie, A., & Bäckström, T. (2022). Voice-Quality Features for Replay Attack Detection. In *2022 30th European Signal Processing Conference (EUSIPCO)* (pp. 384-388). (European Signal Processing Conference). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9909802

# Voice-quality Features for Replay Attack Detection

Abraham Woubie and Tom Bäckström
*Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland*
abraham.zewoudie@aalto.fi, tom.backstrom@aalto.fi

***Abstract*—Replay attacks are mainly carried out to get fraudulent access to an Automatic Speaker Verification (ASV) system. These type of attack requires recording and playback devices. In this paper, we investigate the usefulness of voice-quality features to detect replay attacks. The voice-quality features are used together with the state-of-the-art constant Q cepstral coefficients (CQCC) features. The two feature sets are fused at the score level. Thus, the log-likelihood scores estimated from the two feature sets are linearly weighted to obtain a single fused score. The fused score is used to classify whether a given speech sample is genuine or spoofed. Our experiments with the ASVspoof 2017 dataset demonstrate that the fusion of log-likelihood scores extracted from the CQCC and voice quality features provide better Equal Error Rate (EER) compared to the baseline system which is based only on CQCC features.**

***Index Terms*—fusion, jitter, replay attack, shimmer, spoofing**

## I. Introduction

Automatic speaker verification is widely used in a range of applications which require not only robustness to changes in the acoustic environment, but also resilience to intentional circumvention, known as spoofing [1]. Spoofing is an attack where a fraudster tries to gain access of the system by masquerading as an enrolled person in the automatic speaker verification (ASV) system [1], [2]. The state-of-the-art ASV system are susceptible to different types of spoofing attacks such as voice conversion, speech synthesis and replay attacks [1], [3].

Replay attacks are mounted using recordings of a target speaker's voice which are replayed to an ASV system in the place of genuine speech. Thus, replay attacks are easy to perform and their threat to the reliability of ASV has been studied widely [4]–[6]. Replay attacks use recordings of a target speaker's voice which is replayed to the ASV system in place of genuine speech [7], [8]. A prime example is to record and replay a target speaker's voice to unlock a smartphone which uses ASV for access control.

Recently, replay attacks attracted a lot of attention in the research community. For example, the ASVspoof 2017 challenge provided a standard corpora for combating replay attack [9]. The challenge uses the constant Q cepstral coefficients (CQCC) as the feature set and Gaussian mixture model (GMM) techniques as a classifier [9], [10]. In fact, most state-of-the art of anti-spoofing systems, including [9], use CQCC as a feature set.

After the release of the the ASVspoof 2017 challenge, different types of features have been proposed by different researchers to improve the performance of countermeasures for replay attacks. For instance, features such as spectral peak mapping filter cepstral coefficients, subband spectral centroid magnitude coefficients (SCMC), subband spectral centroid frequency coefficients (SCFC), Teager energy profiles and others have also been used to detect replay attacks [11]–[15].

Jitter and shimmer voice-quality measurements are long-term estimates that discern variations of fundamental frequency and amplitude, respectively. Studies show that these measurements can be used to detect voice pathologies [16], speaking styles and emotions [17], and also identify age and gender [18]. For example, fusing jitter and shimmer voice-quality measurements with the baseline cepstral features improve the performance of Gaussian mixture model (GMM) based speaker recognition systems [19]. Moreover, using jitter and shimmer measurements together with cepstral ones improves the classification accuracy of different speaking styles [17]. Such voice-quality features are also important in speaker diarization [20], and they can be used to characterize different types of voices such as breathy, tense, harsh, whispery and creaky [16].

Thus, the main contribution of this work is that we propose the use of voice-quality features for anti-spoofing systems, particularly to detect replay attacks. The voice-quality features are used together with the state-of-the-art constant Q cepstral coefficients (CQCCs) features. The voice-quality features are fused with the CQCCs at the score likelihood level (i.e., the log-likelihood scores extracted using CQCC and voice-quality models are linearly weighted). We are interested in voice-quality features since jitter and shimmer measurements show significant differences between different speaking styles. In addition, since these features have shown potential for characterizing pathological voices and linguistic abnormalities, they can be also employed to characterize a particular speaker.

## II. Voice-quality Features

Voice-quality features characterize the glottal excitation of signal of voiced voices such as glottal pulse shape and fundamental frequency, and carry speaker-specific information. Analysis of the voice-quality of a person is a valuable technique for speech pathology detection [21], [22]. For example, voice-disorders can be analyzed using such acoustic signal parameters. Unlike the $F_0$, voice-quality features do not always have an acoustic characteristic that is easily distinguishable and measurable from a speech signal.

Jitter and shimmer voice quality features measure variations of the fundamental frequency and amplitude of pitch periods, respectively. They are very useful to describe the fluctuations of the voice signal in a qualitative way. They are given as

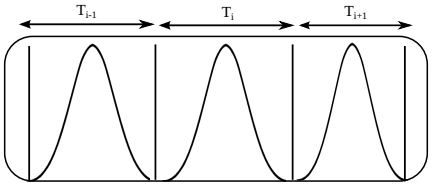



Fig. 1. *Jitter measurements for* 3 *pitch periods*

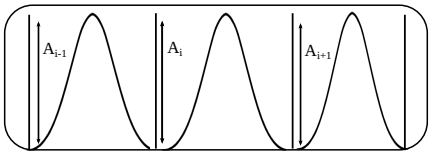



Fig. 2. *Shimmer measurements for* 3 *pitch periods*

a percentage that represents the maximum deviation from a normal frequency or amplitude. There are many possible jitter and shimmer measurements, but usually it is based on an auto-correlation method for determining the frequency and location of each cycle of vibration of the vocal folds (i.e., pitch marks) [23].

Jitter and shimmer voice quality features can be used to detect voice pathologies [24]. They are normally measured from long sustained vowels where voice-quality measurement values outside a certain threshold are considered as pathological voices. In addition, voice quality features are related to the shape and dimension of the speaker's vocal tract, and the way how the speech is generated by the voice production mechanism. Jitter and shimmer can also be used to characterize the age and the gender of a speaker [25]. Moreover, there are also significant differences in jitter and shimmer measurements between different speaking styles, especially in shimmer measurements [26]. Since these features have shown potential for characterizing pathological voices and linguistic abnormalities, they can be also employed to characterize a particular speaker.

There are many possible jitter and shimmer measurements. By using Praat [27], one can extract 5 different jitter and 6 different shimmer measurements.

Although there are different types of jitter and shimmer measurements as it is explained above, we have extracted 5 different jitter and 4 different shimmer measurements encouraged by previous work of [19]. This work reported that these measurements provide better results for speaker recognition more than the other jitter and shimmer measurements. We have extracted the following types of measurements: Jitter (local), Jitter (local, absolute), Jitter (rap), Jitter (ppq5), Shimmer (local), Shimmer (local, dB), Shimmer (apq3) and Shimmer (apq11). A clear description of these measurements is found in [27].

The voice-quality features are used with the state-of-the-art anti-replya attack features (i.e., CQCC). CQCC features are derived using the constant Q transform (CQT) [28], [29], a perceptually motivated time-frequency analysis tool and alternative to the short-term Fourier transform (STFT). Whereas the STFT operates with a fixed spectrotemporal resolution, that of the CQT is variable, with a higher frequency resolution at lower frequencies and a higher temporal resolution at higher frequencies. Liike Melfrequency cepstral coefficients, the CQCC extraction is performed with a filterbank, where the Q factor is a measure of the selectivity of each filter, defined as the ratio between the centre frequency of the filter and its bandwidth [9].

## III. Proposed System

To improve the performance of the baseline anti-spoofing system, we propose a score-level framework that fuses the information provided by CQCC and voice-quality features as it is shown in Fig. III.

Firstly, the training data is partitioned into two sets: genuine and spoofed. Then, two types of features (i.e., CQCC and voice-quality) are extracted from the genuine and spoofed data. Afterwards, we train two types of GMM models using genuine data: one GMM model using CQCC and another GMM model using voice-quality features. Similarly, we train two types of GMM models using spoofed data: one GMM model using CQCC and another GMM model using voice-quality features. Then, the log-likelihoods are predicted using the respective trained models for each feature set. Thus, four different log-likelihoods are computed using the genuine and spoofed models for CQCC and voice-quality features. Finally, the log-likelihoods predicted for the voice-quality are fused together with the log-likelihoods predicted using CQCC features. The GMM models are learned using expectation maximisation (EM) algorithm with random initialisation. Note that the baseline system uses only CQCC feature set.

Given an unseen test utterance, the CQCC and voice-quality features are first computed. Then, they are scored with their respective models to obtain the log-likelihood scores. Afterwards, the two log-likelihood scores predicted using the two models are combined in a weighted fashion such that their weights sum to 1. Finally, the combined scores are used to make a decision (i.e., accept/reject a speech as genuine or spoofed).

The fused cosine-distance score is calculated as follows:

$$\begin{aligned}\Lambda(X) &= \alpha\ (log\ L(X|\Theta_n) - log\ L(X|\Theta_s)) + \\ &= ((1-\alpha)\ (log\ L(X|\vartheta_n) - log\ L(X|\vartheta_s))),\end{aligned} \tag{1}$$

where $\Lambda(X)$ is the fused log-likelihood score, $\Theta_n$ and $\Theta_s$ are the genuine and spoofed GMM model using CQCC features, respectively and $\vartheta_n$ and $\vartheta_n$ are the genuine and spoofed GMM model using voice-quality features, respectively. In addition, two different weights are applied on the predicted log-likelihood scores. While $\alpha$ weights the log-likelihood score predicted using CQCC, $(1-\alpha)$ weights the log-likelihood scores from the voice-quality features.

## IV. Experiments

### A. Experimental Setup

This work used the ASV Spoof 2017 challenge version 2.0 database [9]. The ASV Spoof 2017 database was collected in

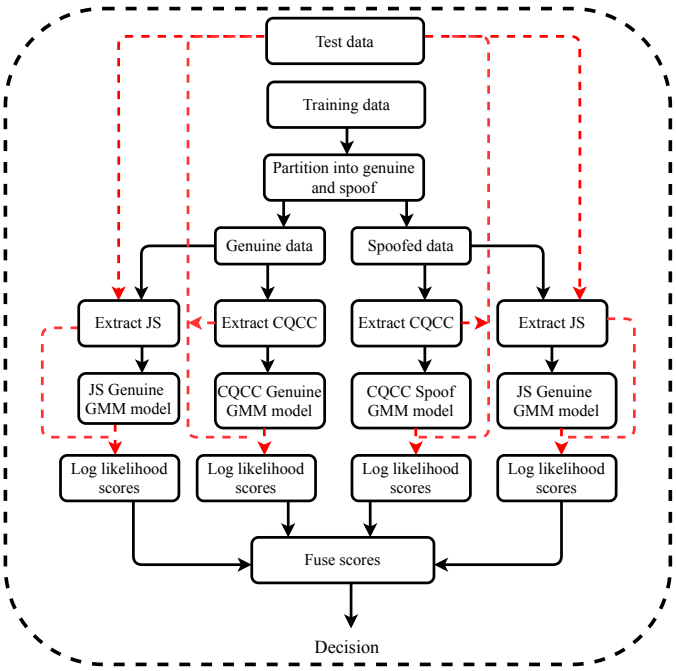



Fig. 3. The proposed replay attack detection system using CQCC, and Jitter and Shimmer (JS) features. The baseline system is based only on CQCC features. While the arrows in black (undotted) correspond to training phase, the arrows in red (dotted) correspond to evaluation.

order to foster the development of countermeasures to protect ASV systems from replay spoofing attacks. It is partitioned into three subsets: training, development and evaluation. The number of files in the training, development and evaluation set are 3014, 1710 and 13306, respectively.

The baseline system [9] uses the CQCC feature set. The maximum and minimum frequency values are set to 8 and 15, respectively. The number of bins per octave is 96 and the total number of CQCC features extracted include 19 static coefficients, log-energy, deltas, and delta-delta. Thus, the CQCC has a feature vector of length 60. After the extraction of CQCC features, the means and variances are normalized. The voice-quality features are extracted over 30 frame length and at 10 shift using Praat [27]. Each of the voice-quality features are then estimated over a 500 window with 10 shift. This is done to smooth out the feature estimation of the unvoiced frames. It is also done to synchronize the voice-quality features with CQCC. We analyzed the smoothing using different window sizes (i.e., 100, 200, 300, 400 and 500ms) on the development set. We have used 500 ms as a smoothing window since it provides us the lowest percentage of zeros values for the unvoiced frames of voice-quality features in the development set.

Both the baseline and the proposed system uses GMM classifier for modeling the classes corresponding to natural and spoofed speech utterances. Since the size of CQCC feature is 60, we set its GMM components to 512. But, since the voice-quality features has only 9 features, its GMM components is set to 128.

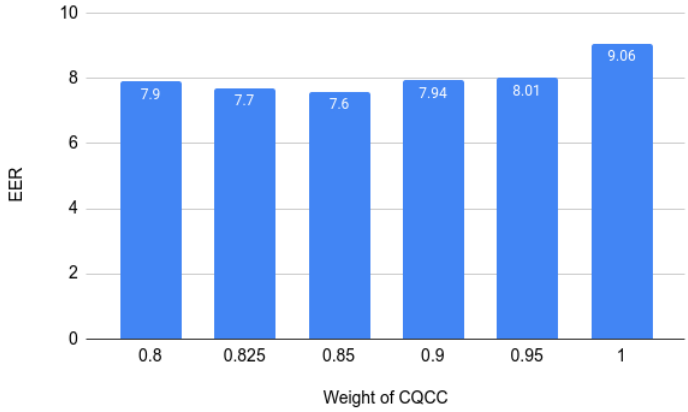



Fig. 4. Equal Error Rate (EER) of the development set when the weight of CQCC is tuned. The baseline system (i.e., cqcc) has weight value of 1.

### B. Experimental Results

We chose to use the work of [9] as a baseline system. We further selected the best results from [9] to compare the performance of the proposed system. In addition to the work of [9], we have also compared our baseline EER with the recent work of [15].

The performance of the proposed and baseline systems were compared with three different experiments. The first system uses the training dataset (3710 files) as a training set and uses two test sets: development and evaluation. The second system uses the development set (1710 files) as training, and uses two test sets: training and evaluation. Finally, the training and development sets are pooled together (i.e., 4724 files) and are used as a training test, and the test is carried out on the evaluation set.

To find the best weight values for CQCC and voice-quality features, we experimented with different values for the weight on the development set as it is shown in Fig. 4. The figure shows that the fusion of voice-quality features with CQCC using different weight values provides better EER compared to the baseline system. Since the best EER value is found when the weight of CQCC is 0.85, we have used weight of 0.85 for CQCC and 0.15 for voice quality features for the results reported in Table I.

From the results of Table I, we see that the baseline system which uses the training data as a training set provides an EER of $9.06\%$ and $13.74\%$ on the development and evaluation sets, respectively. The table further shows that the fusion of voice-quality features with CQCC gives an EER of $7.6\%$ and $12.3\%$ on the development and evaluation sets, respectively. Thus, the results show that the addition of voice-quality measurements to the CQCC feature set provides a $16\%$ and $10.48\%$ relative EER improvement on the development and evaluation sets, respectively.

Similarly, Table I shows that when the development set is used as a training set, the baseline system provides an EER of $5.66\%$ and $14.77\%$ on the training and evaluation sets,

TABLE I

REPLAY DETECTION PERFORMANCE IN TERMS OF EQUAL ERROR RATE (EER) FOR THE ASVSPOOF 2017 VERSION 2.0 DATABASE FOR TRAINING (T), DEVELOPMENT (D) AND TESTING (T) CONFIGURATIONS. JS REPRESENTS JITTER AND SHIMMER VOICE-QUALITY MEASUREMENTS. NOTE THAT THE BASELINE SYSTEM EER (I.E, CQCC) RESULTS ARE TAKEN FROM THE WORK OF [9].

| | **Training** | | | | |
|---|---|---|---|---|---|
| | **T** | | **D** | | **T + D** |
| **Evaluation** | **D** | **E** | **T** | **E** | **E** |
| CQCC [9] | 9.06 | 13.74 | 5.66 | 14.77 | 12.24 |
| CQCC + JS | 7.6 | 12.3 | 4.01 | 13.2 | 10.5 |

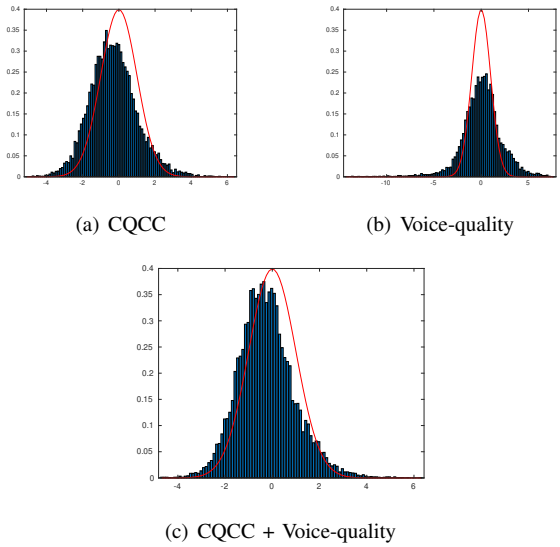
respectively. The table further reveals that the augmentation of voice-quality features with CQCC on the same dataset provides EER of $4.01\%$ and $13.2\%$ on the training and evaluation sets, respectively. These improvements represent a $29.15\%$ and $10.62\%$ relative EER improvement on the training and evaluation sets, respectively.

In a final experiment, we pooled the training and development files together and used them as a training set of 4724 files. The table shows that the baseline system provides an EER of $12.24\%$. However, the addition of voice-quality features to the CQCC reduces the EER to $10.5\%$. This represents a $14\%$ relative EER improvement compared to the baseline system.

The histogram plots of log-likelihood scores obtained from Gaussian mixtures corresponding to (a) CQCC, (b) voice-quality features and (c) CQCC + voice-quality features are shown in Fig. 5. The log-likelihood scores are for the evaluation set. From the figure, we see that the log-likelihood scores of the voice-quality scores of both natural and replay are distributed more resulting in lower % EER as compared to the distribution obtained from CQCC. When the log-likelihood scores of CQCC and voice-quality are fused, the scores of CQCC and voice-quality are multiplied by 0.85 and 0.15, respectively.

In addition, the work in [9] also reported results of replay attack detection without mean variance normalization and without log energy. Thus, in order to further assess the impact of voice-quality features, we have also made another set of experiments where the log-energy is not used at all and no normalization is carried out. Similar to the results reported in Table I, our experimental results showed that the fusion of voice-quality features with the CQCC, irrespective of log-energy and normalization, always provides better EER than the baseline system which is based only on CQCC features, respectively.

Note that in addition to the score level fusion, we have also carried out another experiment by fusing the cqcc with the voice-quality features at the feature level to compare the results of feature fusion with score fusion technique. Thus, we fused the cqcc and voice-quality features to form a 60 X 9 vector and trained a single GMM classifier. However, when we combine the two streams at the feature level, the EER is almost similar to the baseline. Thus, our experimental results show that score level fusion is a better fusion technique for anti-replay detection.



(a) CQCC (b) Voice-quality

(c) CQCC + Voice-quality

Fig. 5. Histogram of log-likelihoods scores of CQCC, voice-quality, and CQCC + Voice-quality. The weights of CQCC and voice-quality features are 0.85 and 0.15 in (c), respectively.

Thus, the results reported in Table I demonstrate that the long-term voice-quality features provide useful and complementary speaker information. The experimental results show that adding jitter and shimmer voice quality features to the baseline CQCC features reduce the EER values. In overall, the use of voice-quality features together with CQCC ones increase the robustness and reliability of anti-spoofing systems. Note that this works analyzed the robustness of voice-quality to spoofing, not background noise.

## V. CONCLUSIONS

In this work, we have proposed the use of jitter and shimmer voice-quality measurements as a complementary source of information to detect replay attacks. The experimental results carried out on ASVspoof 2017 database demonstrate that the fusion of the log-likelihood scores of voice-quality with the log-likelihood scores of CQCC improves the performance of anti replay attacks. The experimental results show that the augmentation of voice-quality features with CQCC provide a 14% relative EER improvement compared to using CQCC features on the evaluation set of ASVspoof 2017 database. Thus, the results reported in this work demonstrate the usefulness of voice-quality measurements as a complementary source of information to detect replay attacks.

The future work could focus on applying deep neural network techniques on these long-term voice-quality measurements to reduce the EER of anti-spoofing systems.

## VI. ACKNOWLEDGMENT

This work has been supported by the Jane and Aatos Erkko foundation funding under contract 700795 AUTHSPKR.

## REFERENCES


[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

[2] H. A. Patil and M. R. Kamble, "A survey on replay attack detection for automatic speaker verification (asv) system," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1047–1053.

[3] N. Evans, J. Yamagishi, and T. Kinnunen, "Spoofing and countermeasures for speaker verification: a need for standard corpora, protocols and metrics," *IEEE Signal Processing Society Speech and Language Technical Committee Newsletter*, pp. 2013–05, 2013.

[4] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2014, pp. 1–6.

[5] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–5.

[6] J. Gałka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, pp. 143–153, 2015.

[7] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification-a study of technical impostor techniques," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[8] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, 2010, pp. 131–134.

[9] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements," 2018.

[10] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

[11] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 1678–1681.

[12] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the asvspoof 2017 challenge." in *Interspeech*, 2017, pp. 7–11.

[13] M. R. Kamble and H. A. Patil, "Analysis of reverberation via teager energy features for replay spoof speech detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2607–2611.

[14] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection." in *Interspeech*, 2018, pp. 641–645.

[15] M. R. Kamble and H. A. Patil, "Novel variable length teager energy profiles for replay spoof detection," *energy*, vol. 32, p. 33, 2020.

[16] J. Kreiman and B. R. Gerratt, "Perception of aperiodicity in pathological voice," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2201–2211, 2005.

[17] 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 *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–1081.

[18] A. S. Naini and M. Homayounpour, "Speaker age interval and sex identification based on jitters, shimmers and mean mfcc using supervised and unsupervised discriminative classification methods," in *2006 8th international Conference on Signal Processing*, vol. 1. IEEE, 2006.

[19] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Eighth annual conference of the international speech communication association*, 2007.

[20] A. Woubie, J. Luque, and J. Hernando, "Using voice-quality measurements with prosodic and spectral features for speaker diarization," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, and G. S. Berke, "Comparison of voice analysis systems for perturbation measurement," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 1, pp. 126–134, 1996.

[22] I. C. Zwetsch, R. D. R. Fagundes, T. Russomano, and D. Scolari, "Digital signal processing in the differential diagnosis of benign larynx diseases," *Scientia Medica*, vol. 16, no. 3, pp. 109–114, 2006.

[23] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease," *The journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[24] I. Wagner, "A new jitter-algorithm to quantify hoarseness: an exploratory study," *International Journal of Speech Language and the Law*, vol. 2, no. 1, pp. 18–27, 2013.

[25] N. Minematsu, M. Sekiguchi, and K. Hirose, "Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–137.

[26] 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 *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1081.

[27] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2009.

[28] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[29] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.