Voice-quality Features for Deep Neural Network Based Speaker Verification Systems

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

***Abstract*—Jitter and shimmer are voice-quality features which have been successfully used to detect voice pathologies and classify different speaking styles. In this paper, we investigate the usefulness of such voice-quality features in neural-network based speaker verification systems. To combine these two sets of features, the cosine distance scores estimated from the two sets are linearly weighted to obtain a single, fused score. The fused score is used to accept/reject a given speaker. The experimental results carried out on Voxceleb-1 dataset demonstrate that the fusion of the cosine distance scores extracted from the mel-spectrogram and voice quality features provide a 15% relative improvement in Equal Error Rate (EER) compared to the baseline system which is based only on mel-spectrogram features.**

***Index Terms*—jitter, mel-spectrogram, fusion, shimmer, speaker recognition**

I. Introduction

Speech analysis methods relies on effective feature extraction, which is used to retrieve relevant information from the acoustic signal. The feature extraction module therefore needs to extract features that have large between-speaker variability and small within-speaker variability. Most of the state-of-the-art speaker verification systems use only the short-term features such as MFCC or the mel-spectrogram [1], [2].

While short-term features capture the local speech characteristics in a short time window, long-term features reflect voice characteristics over a whole utterance. Thus, long-term features capture phonetic, prosodic, lexical, syntactic, semantic and pragmatic information. Short-term features are extracted from a single speech frame, while long-term features are extracted from portions of speech longer than one frame. Since long-term features provide discriminative power, fusion of short-term spectral features with long-term features has been applied on different speech applications [3]–[5]. Long-term speech features are also robust to channel variation since temporal patterns do not change with the change of acoustic conditions [6].

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 [7], speaking styles and emotions [8], and also identify age and gender [9]. 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 [10]. Moreover, using jitter and shimmer measurements together with cepstral ones improves the classification accuracy of different speaking styles [8]. Such voice-quality features are also important in speaker diarization [5], [11]–[13], and they can be used to characterize different types of voices such as breathy, tense, harsh, whispery and creaky [7].

The main contribution of this work is that we propose the use of voice-quality features for deep learning based speaker verification systems. The voice-quality features are used together with the short-term mel-spectrogram features. The fusion of the voice-quality features with the mel-spectrogram is carried out at the score likelihood level, i.e., the cosine distance scores extracted using the mel-spectrogram 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. Since these features have shown potential for characterizing pathological voices and linguistic abnormalities, they can be also employed to characterize a particular speaker.

The rest of this paper is organized as follows. The next section gives an overview of voice-quality features used in our work. Section III described the architecture of the proposed system and the fusion technique. Experimental results and conclusions are presented in Section IV and Section V, respectively.

II. Voice-quality Features

Voice-quality features characterize the glottal excitation 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 [14]. Voice quality is composed of many aspects of the speech production. It is characterized by qualitative terms such as hoarseness, whispering, creakiness, etc. The acoustic parameters can be used to detect if a person has a pathological problem. The most widely used acoustic parameters used to assess the quality of a voice are jitter, shimmer and harmonics-to-noise ratio.

The calculation of jitter and shimmer measurements is usually based on an autocorrelation method for determining the frequency and location of each cycle of vibration of the

vocal folds (i.e., pitch marks) [15]. In addition to this, 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.

There are many possible jitter and shimmer measurements. By using Praat [16], one can extract 5 different jitter and 6 different shimmer measurements. In this work, we have extracted 4 jitter and 5 shimmer measurements.

A. Jitter

Deviations from the mean pitch period length of a voice signal are known as jitter. Ideally, each cycle of a speech signal would have the same period length. Jitter measures how much one period differs from the next in the speech signal. It is mainly due to fluctuations in the opening and closing times of the vocal folds and they introduce a distortion which appears as a frequency modulation in the speech signal. Jitter is a useful measure in speech pathology since pathological voices often have a higher jitter than healthy voices [17]. The values of jitter can be higher because of a number of conditions that affect the vocal cords such as nodules, polyps, and weakness of the laryngeal muscles. We extract the following types of jitter measurements:

- **Jitter (local)**: The average absolute difference between consecutive period lengths, divided by the average period length.
- **Jitter (local, absolute)**: The average absolute difference between consecutive period lengths in seconds.
- **Jitter (rap)**: The relative average perturbation is the average absolute difference between a period and the average of it and its two neighbours, divided by the average period.
- **Jitter (ppq5)**: The five-point period perturbation quotient is the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period.

B. Shimmer

Similar to jitter, but instead of looking at periodicity, shimmer quantifies the difference in amplitude from cycle to cycle. Shimmer changes with the reduction of glottal resistance and mass lesions on the vocal cords and is correlated with the presence of noise emission and breathiness. It is also a useful measurement in speech pathology since pathological voices often have higher shimmers values more than the healthy voices [17].

We extract the following types of shimmer measurements:

- **Shimmer (local)**: The average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.
- **Shimmer (local, dB)**: The average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20.
- **Shimmer (apq3)**: The three-point amplitude perturbation quotient is the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude.
- **Shimmer (apq5)**: The five-point amplitude perturbation quotient is the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude.
- **Shimmer (apq11)**: The 11-point amplitude perturbation quotient is the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its ten closest neighbours, divided by the average amplitude.

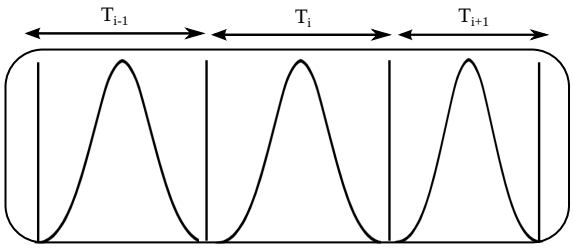

Fig. 1. *Jitter measurements for* 3 *pitch periods*

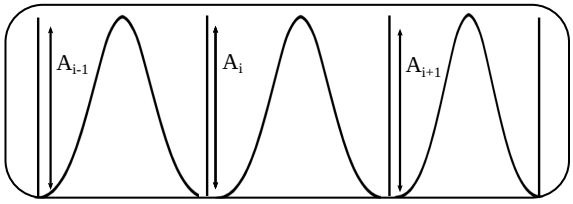

Fig. 2. *Shimmer measurements for* 3 *pitch periods*

III. Proposed Architecture and Fusion Technique

To improve the performance of the CNN based speaker verification system, we propose a score-level framework that fuses the information provided by mel-spectrograms and voice-quality features. Let the number of speakers to be enrolled for the speaker verification system be N. The enrollment data is used to train two sets of convolutional neural network (CNN) models: one model using the mel-spectrogram and another model using voice-quality features. Given an unseen test utterance, the mel-spectrogram and voice-quality features are first computed. Then, they are scored with their respective models to obtain two sets of cosine distance scores. Afterwards, the two cosine distance scores predicted using the two models are combined in a weighted fashion such that the weights sum to 1. Finally, the combined scores are used to make a decision (i.e., accept/reject a speaker identity).

Firstly, jitter and shimmer voice quality features are extracted from the fundamental frequency contour. Then, they are fused together with the baseline mel-spectrogram features. The fusion of the two streams is carried out at the score likelihood level, i.e., we combine the cosine distance scores predicted

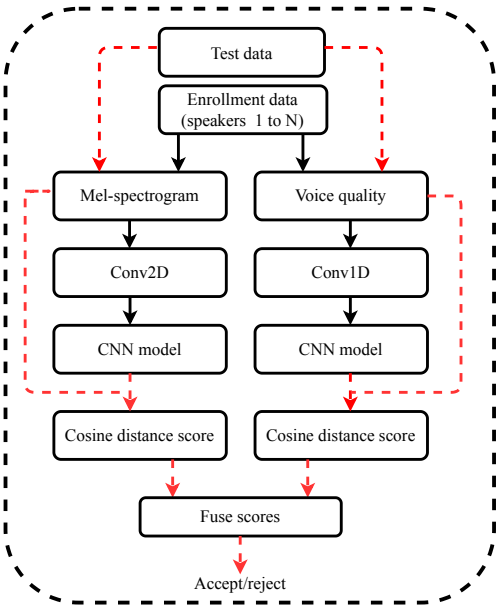

Fig. 3. The proposed CNN based speaker verification system using short-term mel-spectrogram and long-term voice-quality features. While the arrows in black (undotted) correspond to training (enrollment) phase, the arrows in red (dotted) correspond to evaluation.

using the mel spectrogram model and voice-quality feature trained model.

The fused cosine-distance score is

$$\text{score}(i,j) = \beta \frac{\mathbf{x_i}^T \mathbf{x_j}}{\|\mathbf{x_i}\|\|\mathbf{x_j}\|} + (1-\beta)\frac{\mathbf{y_i}^T \mathbf{y_j}}{\|\mathbf{y_i}\|\|\mathbf{y_j}\|}, \quad (1)$$

where the scalar score(i,j) is the fused cosine distance score of test file i and test file j, $\mathbf{x_i}$ and $\mathbf{x_j}$ are the corresponding speaker embeddings extracted using the mel-spectrogram CNN model for test file i and test file j, respectively and $\mathbf{y_i}$ and $\mathbf{y_j}$ are the speaker embeddings extracted using the voice-quality CNN model for the same test files i and j, respectively. In addition, two different weights are applied on the predicted cosine-distance scores. While β weights the cosine-distance predicted using speaker embeddings extracted using mel-spectrogram CNN trained model, $(1-\beta)$ weights the cosine-distance of speaker embeddings extracted from the voice-quality features CNN model.

IV. EXPERIMENTS

A. Database and experimental setup

The input features of the baseline system are mel-spectrograms, computed within a 30ms frame window at 10ms shift using Librosa [18]. Mean and variance normalization is performed on every frequency bin of the spectrum. The voice-quality features are extracted over 30ms frame length and at 10ms shift using Praat software [16]. Each of the voice-quality features are then estimated over a 500 ms window with 10ms 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 the short-term ones. 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 mel-spectrograms and voice-quality features are extracted from the first 3.5 seconds of Voxceleb-1 audio files. Thus, the sizes of the mel-spectrogram and voice-quality features are 350×80, and 350×9, respectively. Since the size of mel-spectrograms is 350 by 80, we use "Conv 2D". But, we use "Conv 1D" for the voice-quality features have size of 350 by 9.

Our system was implemented using the Keras deep learning library [19] to train the two models: one model using mel-spectrogram and another model using voice-quality features. Each network is trained on a Titan X GPUs for 100 epochs or until the validation error stops decreasing, whichever is sooner, using a batch-size of 64. We use SGD with momentum (0.9), weight decay ($5E-4$) and a logarithmically decaying learning rate (initialised to 10^{-2} and decaying to 10^{-8}).

The proposed speaker verification system has been carried out on the VoxCeleb-1 database [20]. It contains 148,642 development and 4,874 test utterances, which belong to 1211 and 40 speakers, respectively. From the test set, 37,720 experimental trials were scored. Half of them are client trials while the other half are impostor trials.

Performance was evaluated using two performance metrics: (i) the Equal Error Rate (EER) which is the rate at which both acceptance and rejection errors are equal; and (ii) the cost function

$$C_{det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}) \quad (2)$$

where we assume a prior target probability P_{tar} of 0.01 and equal weights of 1.0 between misses C_{miss} and false alarms C_{fa}. Both metrics are commonly used for evaluating identity verification systems.

Note that in order to validate the generalization of results, the weight parameters in Eq. 1 were first tuned using few data from the development set of Voxceleb-1. Then, the tuned weight values have been directly used on the Voxceleb-1 test set. In the development set, a weight value of 0.9 and 0.1 gave us the best EER values for the spectrogram and voice-quality features, respectively. Thus, in the test set, we evaluated the EER using a weight value of 0.9 for the spectrogram and 0.1 for the voice-quality features.

B. Experimental results

Figure 4 shows that the baseline system which is based only on mel-spectrogram features has an EER of 8.1%. Our baseline EER on Voxceleb-1 dataset is almost similar to other similar works that use CNN architecture for speaker verification [20], [21].

The figure shows that the fusion of the mel-spectrogram with the 9 voice-quality features reduces the EER to 6.9%.

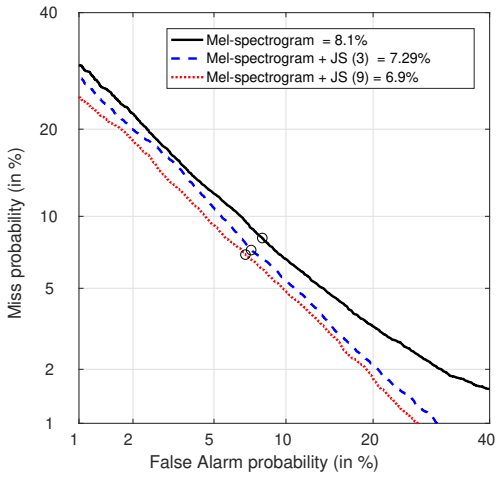

Fig. 4. EER of the baseline and proposed system. While the baseline system is based only on mel-spectrogram features, the proposed system uses mel-spectrogram together with voice-quality measurements. JS (3) and JS (9) represent the use of three and nine types of jitter and shimmer measurements, respectively

This represents a 14.8% relative EER improvement compared to the baseline system. Encouraged by the previous works of [22], we have also carried out another experiment where we use only absolute jitter, absolute shimmer and shimmer apq3 measurements since these three measurements proved to be useful for speaker diarization. Table I shows that the fusion of mel-spectrogram with these three voice-quality measurements provide an EER of 7.29%, which is almost a 10% relative EER improvement compared to the baseline system.

In addition to EER, we have also compared the minimum detection cost function (minDCF) values of the baseline and proposed system. As it is reported in Table II, the minDCF value of the baseline system is 0.72. While the fusion of the mel-spectrogram with the three voice-quality features reduce the minDCF value to 0.7, the fusion of the mel-spectrogram with the nine voice-quality features reduce the minDCF value to 0.57.

Thus, the results reported in Figure 3 and Table II demonstrate that the voice-quality features provide useful and complementary speaker information. The experimental results show that adding jitter and shimmer voice quality features to the baseline mel-spectrogram features reduce both the EER and minDCF values.

In order to generalize the results reported in Table I, we have also analyzed the EER and minDCF values of the baseline and proposed system by partitioning the Voxceleb-1 37,720 test trials into 4 equal partitions. Thus, each partition has 9430 trial files. The results reported in Table II demonstrate that both EER and minDCF values of the proposed system for the whole partition sets are better than the baselines system which uses only mel-spectrogram features. Thus, the results of Table I and Table II show the usefulness of voice-quality features for deep neural network based speaker verification

TABLE I
EER AND MINDCF OF THE BASELINE AND PROPOSED SYSTEM. JS (3) AND JS (9) REPRESENT THE USE OF THREE AND NINE TYPES OF JITTER AND SHIMMER MEASUREMENTS, RESPECTIVELY.

Features	**EER**	**minDCF**
Mel-spectrogram (Baseline)	8.1%	0.72
Mel-spectrogram + JS (3)	7.29%	0.7
Mel-spectrogram + JS (9)	**6.9%**	**0.57**

TABLE II
EER AND MINDCF OF THE BASELINE AND PROPOSED SYSTEM AFTER PARTITIONING THE VOXCELEB-1 37,720 TEST TRIALS INTO 4 EQUAL PARTITIONS (I.E., EACH PARTITION HAS 9430 TRIAL FILES).

	Features			
	Mel-spectrogram		Mel-spectrogram + Voice-quality	
	EER(%)	minDCF	EER(%)	minDCF
Partition 1	10.73	0.7	8.37	0.66
Partition 2	7.03	0.6	7.05	0.52
Partition 3	6.95	0.6	5.59	0.48
Partition 4	7.3	0.57	6.41	0.4

systems. The experimental results demonstrate that the voice-quality features convey useful and complementary speaker information to the mel-spectrograms.

Note that in addition to the score level fusion, we have also carried out another experiment by fusing the spectrogram with the voice-quality features at the feature level to compare the results of feature fusion with score fusion technique. Thus, we fused the spectrogram and voice-quality features to form a 350 X 9 vector and trained a single CNN. In a preliminary experiment we have conducted to analyze the impact of feature fusion, the feature fusion technique does provide better result than the baseline system. Thus, in the future, in-depth experiments in this direction would be interesting in order to confirm our findings.

V. CONCLUSIONS

In this work, we have proposed the use of jitter and shimmer voice-quality measurements as complementary source of information to CNN based speaker verification system. Experimental results on Voxceleb-1 corpus show that the fusion of the voice-quality with the mel-spectrograms at the score level increases speaker verification performance. The experimental results show that the augmentation of voice-quality features to the mel-spectrogram provide almost a 15% relative EER improvement. Thus, the results reported in this work manifest the usefulness of voice-quality measurements as complementary source of information for neural network based speaker verification system.

The future work could focus on extracting i-vectors from the voice-quality features and analyze their impact using both cosine distance and Probabilistic Linear Discriminant Analysis (PLDA) scoring techniques.

VI. ACKNOWLEDGMENT

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

REFERENCES

[1] D. Snyder, J. Villalba, N. Chen, D. Povey, G. Sell, N. Dehak, and S. Khudanpur, "The jhu speaker recognition system for the voices 2019 challenge." in *INTERSPEECH*, 2019, pp. 2468–2472.

[2] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[3] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, "Prosodic and other long-term features for speaker diarization," *in IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 985–993, 2009.

[4] M. Zelenák and J. Hernando, "The detection of overlapping speech with prosodic features for speaker diarization." in *Interspeech*, 2011, pp. 1041–1044.

[5] 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.

[6] E. Shriberg, "Higher-level features in speaker recognition," in *Speaker Classification I*. Springer, 2007, pp. 241–259.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] A. W. Zewoudie, J. Luque, and F. J. Hernando Pericás, "Jitter and shimmer measurements for speaker diarization," in *VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop: proceedings: November 19-21, 2014: Escuela de Ingeniería en Telecomunicación y Electrónica Universidad de Las Palmas de Gran Canaria: Las Palmas de Gran Canaria, Spain*, 2014, pp. 21–30.

[12] A. Woubie, J. Luque, and J. Hernando, "Improving i-vector and plda based speaker clustering with long-term features." in *INTERSPEECH*, 2016, pp. 372–376.

[13] A. W. Zewoudie, J. Luque, and J. Hernando, "The use of long-term features for gmm-and i-vector-based speaker diarization systems," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–11, 2018.

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

[15] 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.

[16] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 5.3. 82)[computer software]," *Amsterdam: Institute of Phonetic Sciences*, 2012.

[17] W. Styler, "Using praat for linguistic research," *University of Colorado at Boulder Phonetics Lab*, 2013.

[18] B. McFee, M. McVicar, C. Raffel, D. Liang, O. Nieto, J. Moore, D. Ellis, D. Repetto, P. Viktorin, J. F. Santos *et al.*, "Librosa: v0. 4.0," *Zenodo 2015*, 2015.

[19] F. Chollet *et al.*, "Keras (2015)," 2017.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[21] H. Salehghaffari, "Speaker verification using convolutional neural networks," *arXiv preprint arXiv:1803.05427*, 2018.

[22] A. Woubie, J. Luque, and J. Hernando, "Short-and long-term speech features for hybrid hmm-i-vector based speaker diarization system," in *Odyssey 2016-The Speaker and Language Recognition Workshop*, 2016.