Jitter and Shimmer Measurements for Speaker Recognition

Mireia Farrús, Javier Hernando, Pascual Ejarque

TALP Research Center, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain
{mfarrus,javier,pascual}@gps.tsc.upc.edu

Abstract

Jitter and shimmer are measures of the cycle-to-cycle variations of fundamental frequency and amplitude, respectively, which have been largely used for the description of pathological voice quality. Since they characterise some aspects concerning particular voices, it is a priori expected to find differences in the values of jitter and shimmer among speakers. In this paper, several types of jitter and shimmer measurements have been analysed. Experiments performed with the Switchboard-I conversational speech database show that jitter and shimmer measurements give excellent results in speaker verification as complementary features of spectral and prosodic parameters.

Index Terms: speaker recognition, jitter, shimmer, prosody, voice spectrum, fusion

1. Introduction

State-of-the-art speaker recognition systems tend to use only short-term spectral features as voice information. Spectral parameters take into account some aspects of the acoustic level of the signal, like spectral magnitudes, formant frequencies, etc., and they are highly related to the physical traits of the speaker. However, humans tend to use several linguistic levels like lexicon, prosody or phonetics to recognise others with voice. These levels of information are more related to learned habits or style, and they are mainly manifested in the dialect, sociolect or idiolect of the speaker.

Since these linguistic levels play an important role in the human recognition process, a lot of effort has been placed in adding this kind of information to automatic speaker recognition systems. [1] showed that idiolectal information provided a good recognition performance given a sufficient amount of data, and more recent works [2-4] have demonstrated that prosody helps to improve voice spectrum based recognition systems, supplying complementary information not captured in the traditional acoustic systems. Moreover, some of these parameters have the advantage of being more robust to some common problems like noise, transmission channel, speech level or distance between the speaker and the microphone than spectral features.

There are probably many more characteristics which may provide complementary information and should be of a great value for speaker recognition. This work focuses on the use of jitter and shimmer for a speaker verification system. Jitter and shimmer are acoustic characteristics of voice signals, and they are quantified as the cycle-to-cycle variations of fundamental frequency and waveform amplitude, respectively. Both features have been largely used to detect voice pathologies (see, e.g. [5, 6]). They are commonly measured for long sustained vowels, and values of jitter and shimmer above a certain threshold are considered being related to pathological voices, which are usually perceived by humans as breathy, rough or hoarse voices. In [7] it was reported that significant differences can occur in jitter and shimmer measurements between different speaking styles, especially in shimmer measurement. Nevertheless, prosody is also highly-dependant on the emotion of the speaker, and prosodic features are useful in automatic recognition systems even when no emotional state is distinguished.

The aim of this work is to improve a prosodic and voice spectral verification system by introducing new features based on jitter and shimmer measurements. The experiments have been done over the Switchboard-I conversational speech database. Fusion of different features has been performed at the score level by using z-score normalization and matcher weighting fusion method.

This paper is organised as follows. In the next section, an overview of the features used in this work is presented, including a description of jitter and shimmer measurements. The experimental setup and verification experiments are shown in section 3. Finally, conclusions of the experiments are given in section 4.

2. Voice features

Cepstral coefficients are the usual way of representing the short-time spectral envelope of a speech frame in current speaker recognition systems. These parameters are the most prevalent representations of the speech signal and contain a high degree of speaker specificity. However, cepstral coefficients have some disadvantages that are overcome by using Frequency Filtering (FF) parameters. These parameters have been used in our experiments since they give comparable or better results than mel-cepstrum coefficients in most of the experiments that have been done [8, 9].

Prosodic parameters are known as suprasegmental parameters since the segments affected (syllables, words and phrases) are larger than phonetic units. These features are mainly manifested as sound duration, tone and intensity variation. The prosodic recognition baseline system used in this work is constituted by nine prosodic features already used in [2, 3]: three features related to word and segmental durations and six features related to fundamental frequency, all of them averaged over all words with voiced frames.

The novel component in this paper is the analysis of jitter and shimmer features in order to test their usefulness in speaker verification. These features have been extracted by using the Praat voice analysis software [10]. Praat reports different kinds of measurements for both jitter and shimmer features, which are listed below.

2.1. Jitter measurements

- *Jitter (absolute)* is the cycle-to-cycle variation of fundamental frequency, i.e. the average absolute difference between consecutive periods, expressed as:

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - T_{i+1}\right| \quad (1)$$

where T_i are the extracted F_0 period lengths and N is the number of extracted F_0 periods.

- *Jitter (relative)* is the average absolute difference between consecutive periods, divided by the average period. It is expressed as a percentage:

$$Jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - T_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \quad (2)$$

- *Jitter (rap)* is defined as the Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbours, divided by the average period.
- *Jitter (ppq5)* is the five-point Period Perturbation Quotient, computed as the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period.

2.2. Shimmer measurements

- *Shimmer (dB)* is expressed as the variability of the peak-to-peak amplitude in decibels, i.e. the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20:

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(A_{i+1}/A_i\right)\right| \quad (3)$$

where A_i are the extracted peak-to-peak amplitude data and N is the number of extracted fundamental frequency periods.

- *Shimmer (relative)* is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as a percentage:

$$Shimmer(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - A_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \quad (4)$$

- *Shimmer (apq3)* is the three-point Amplitude Perturbation Quotient, 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)* is defined as the five-point Amplitude Perturbation Quotient, 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)* is expressed as the 11-point Amplitude Perturbation Quotient, 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.

3. Recognition experiments

3.1. Experimental setup

All the recognition experiments described in this paper have been performed with the Switchboard-I database [11], a collection of 2430 two-sided telephone conversations among 543 speakers from all areas of the United States.

In the prosody based recognition system, a nine-feature vector (already used in [2]) was obtained for each conversation side: three features related to word and segmental durations - number of frames per word and length of word-internal voiced and unvoiced segments - and six features related to fundamental frequency - mean, maximum, minimum, range, pseudo-slope and slope -. Another feature vector was extracted for the acoustic system based on the nine jitter and shimmer measurements described in section 2.

Features were extracted using the Praat software for acoustic analysis [10], performing an acoustic periodicity detection based on a cross-correlation method, with a window length of 40/3 ms and a shift of 10/3 ms. The mean and standard deviation over all words were computed for each individual feature. The system was tested using the *k*-Nearest Neighbour classifier (with *k*=3), comparing the distance of the test feature vector to the *k* closest vectors of the claimed speaker vs. the distance of the test vector to the *k* closest vectors of the cohort speakers. The symmetrised Kullback-Leibler divergence expressed as:

$$d_{KL} = \frac{1}{2}\left(\mu_1 - \mu_2\right)^2\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) + \left(\frac{\sigma_1}{\sigma_2} - \frac{\sigma_2}{\sigma_1}\right)^2 \quad (5)$$

where μ is the mean and σ the standard deviation, was used as a distance measure.

The spectrum based recognition system was a 32-component GMM-UBM system using short-term feature vectors consisting of 20 Frequency Filtering parameters [8] with a frame size of 30 ms and a shift of 10 ms. 20 corresponding delta and acceleration coefficients were included, and the UBM was trained with 116 conversation sides.

All the systems used 8 conversation sides to train the speaker models. Training was performed using splits 1-3 of Switchboard-I database. The three held out splits provided the cohort speakers in prosodic and jitter-shimmer based systems. The systems were tested with one conversation-side according to the NIST's 2001 Extended Data task [12]. Fusion of individual features was performed at the score level for splits 1-3, using the matcher weighting method [13] with a previous z-score normalization. Weights were trained from the splits 4-6 using splits 1-3 as cohort speakers.

3.2. Verification results

First of all, the prosodic system used as baseline is presented. Table 1 shows the EER obtained for each individual prosodic feature and the resulting fusion of the prosodic set.

Table 1. *EER for prosodic features (isolated and fused).*

Feature	EER (%)
log (#frames/word)	31.5
length of word-internal voiced segments	30.0
length of word-internal unvoiced segments	30.0
log (mean F_0)	20.3
log (max F_0)	20.9
log (min F_0)	22.3
log (range F_0)	26.6
pseudo-slope: (last F_0 - first F_0)/(#frames)	38.3
F_0 slope	29.9
Fusion	15.8

The same experiments were performed for the jitter and shimmer measurements described in section 2. Tables 2 and 3 show the EER results for jitter and shimmer features respectively. Both tables give the EER for the individual measurements and the combination of the measurements set.

Table 2. *EER for jitter measurements.*

Jitter measurement	EER (%)
Jitter (absolute)	26.9
Jitter (relative)	33.7
Jitter (rap)	34.2
Jitter (ppq5)	33.8
Fusion	29.2

Table 3. *EER for shimmer measurements.*

Shimmer measurement	EER (%)
Shimmer (dB)	26.9
Shimmer (relative)	28.9
Shimmer (apq3)	28.1
Shimmer (apq5)	32.9
Shimmer (apq11)	33.8
Fusion	25.5

The results show that at least both absolute measurements of jitter and shimmer are potentially useful in speaker recognition. In the case of jitter, its relative measurements do not seem to supply helpful information, since the fusion of all jitter measurements does not outperform the result obtained with the isolated absolute measurement. In order to ensure this assumption, the absolute measurement of jitter was fused with the best-performing relative measurement: the *Jitter (relative)*. The combination of both measurements provided an EER of 29.3%, so that fusion of both measurements does not improve the absolute jitter measurement result either.

In the case of shimmer measurements, their final fusion improves slightly the best isolated result (*Shimmer (dB)*). Since all relative measurements of the same feature are highly correlated, we will only use the relative measurement of shimmer giving the best EER: the *Shimmer (apq3).* To ensure that this measurement provides some complementary information to *Shimmer (absolute)*, both measurements were combined. The EER obtained in the fusion equalled 26.3%, improving slightly the isolated absolute measurement of shimmer.

From now on, only three cycle-to-cycle variability measurements will be used as new features: *Jitter (absolute)*, *Shimmer (dB)* and *Shimmer (apq3)*, and we will refer to this set of three measurements as the *JitShim* system. The EER of the combination of these measurements equals 22.5%.

In order to see how jitter and shimmer are able to improve the prosodic and the voice spectral based recognition systems, the new features are added to both systems separately. First of all, the nine prosodic features used in our baseline system are combined with the three features of our novel *JitShim* system, resulting in a new twelve-featured system. Secondly, the *JitShim* system is added to our voice spectral baseline system. This allows comparing how complementary jitter and shimmer are to prosodic and spectral features, respectively. Finally, the *JitShim* system is combined with both baselines, in order to see how the new features improve our speaker verification system. The results of these experiments are shown in Table 4. The EER before the introduction of the *JitShim* system are given in the middle column of the table, and results after adding jitter and shimmer features are shown in the right column.

Table 4. *EER (%) for prosodic and spectral systems before and after adding jitter and shimmer features.*

Baseline system	without JitShim	with JitShim
Prosodic	15.8	13.1
Spectral	10.1	8.6
Fusion	7.7	6.8

The results and the DET curves plotted in Fig.1 show that both prosodic and spectral baselines are clearly improved when jitter and shimmer features are added to the systems. The best relative improvement is achieved by adding the *JitShim* to the prosody based system (17%). By fusing *JitShim* with the spectral system, the improvement is less considerable (15%). That suggests that the information provided by jitter and shimmer to prosodic parameters is more complementary than the information supplied to the spectral system.

Our preliminary speaker verification system based on prosodic and spectral parameters is also improved by adding the *JitShim* system, as in can be seen in the DET curves plotted in Fig. 2, achieving the lowest EER equalling 6.8%. So, jitter and shimmer features seem to be useful in speaker recognition and should be taken into account in future experiments.

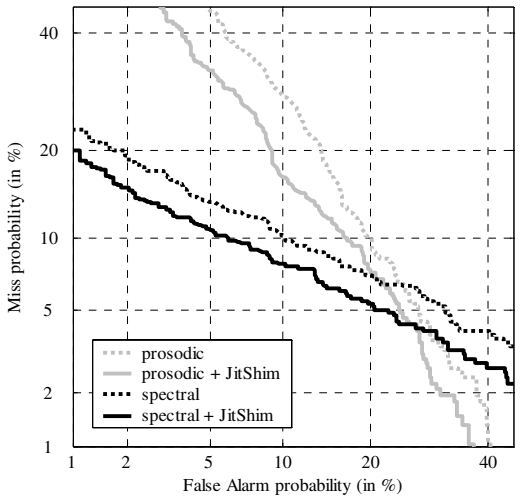

Figure 1. *DET curves for prosodic and spectral systems before and after adding jitter and shimmer features.*

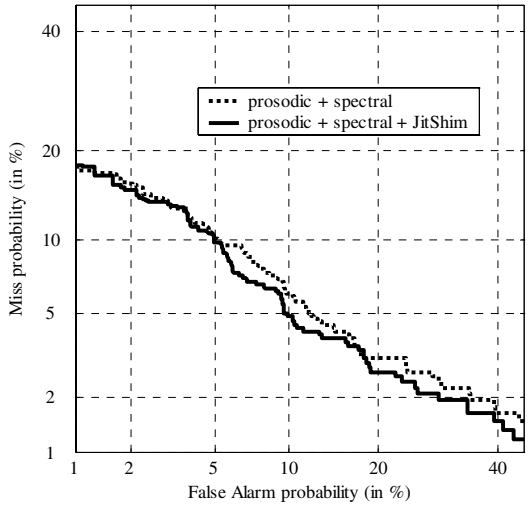

Figure 2. *DET plot showing the improvement of the baseline system after adding jitter and shimmer.*

4. Conclusions

In this work, a preliminary speaker verification system based on prosodic and spectral parameters is improved by adding jitter and shimmer features, which analyse the perturbation of fundamental frequency and waveform amplitude, respectively. In these experiments, the absolute measurements of both features seem to be more discriminant than their relative measurements. Furthermore, the results show that jitter and shimmer can provide complementary information to both spectral and prosodic systems, especially to the prosodic one.

5. Acknowledgements

The authors would like to thank Jan Anguita for his contribution with the voice spectrum based system and Michael Wagner for his valuable comments.

6. References

[1] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," presented at Eurospeech, 2001.

[2] M. Farrús, A. Garde, P. Ejarque, J. Luque, and J. Hernando, "On the Fusion of Prosody, Voice Spectrum and Face Features for Multimodal Person Verification," presented at ICSLP, Pittsburgh, 2006.

[3] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. A. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02," presented at ICASSP, 2003.

[4] D. A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition," presented at ICASSP, 2003.

[5] J. Kreiman and B. R. Gerrat, "Perception of aperiodicity in pathological voice," *Acoustical Society of America*, vol. 117, pp. 2201-2211, 2005.

[6] D. Michaelis, M. Fröhlich, H. W. Strube, E. Kruse, B. Story, and I. R. Titze, "Some simulations concerning jitter and shimmer measurement," presented at 3rd International Workshop on Advances in Quantitative Laryngoscopy, Aachen, Germany, 1998.

[7] R. E. Slyh, W. T. Nelson, and E. G. Hansen, "Analysis of mrate, shimmer, jitter, and F0 contour features across stress and speaking style in the SUSAS database," presented at ICASSP, 1999.

[8] C. Nadeu, J. Hernando, and M. Gorricho, "On the decorrelation of filter bank energies in speech recognition," presented at Eurospeech, 1995.

[9] A. Abad, C. Nadeu, J. Hernando, and J. Padrell, "Jacobian Adaptation based on the Frequency-Filtered Spectral Energies," presented at Eurospeech, Geneva, Switzerland, 2003.

[10] Praat software website: http://www.fon.hum.uva.nl/praat/.

[11] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," presented at ICASSP, 1990.

[12] NIST 2001 Speaker Recognition Evaluation website: http://www.nist.gov/speech/tests/spk/2001/index.htm.

[13] M. Indovina, U. Uludag, R. Snelik, A. Mink, and A. Jain, "Multimodal Biometric Authentication Methods: A COTS Approach," presented at MMUA, Workshop on Multimodal User Authentication, Santa Barbara, CA, 2003.