USING JITTER AND SHIMMER IN SPEAKER VERIFICATION

Mireia Farrús, Javier Hernando

TALP Research Centre, Department of Signal Theory and Communications

Universitat Politècnica de Catalunya, C/ Jordi-Girona 1-3, 08034, Barcelona, Spain

{mfarrus, javier}@gps.tsc.upc.edu

Abstract

Jitter and shimmer are measures of the fundamental frequency and amplitude cycle-to-cycle variations, respectively. Both features have been largely used for the description of pathological voices, and since they characterise some aspects concerning particular voices, they are expected to have a certain degree of speaker specificity. In the current work, jitter and shimmer are successfully used in a speaker verification experiment. Moreover, both measures are combined with spectral and prosodic features using several types of normalisation and fusion techniques in order to obtain better verification results. The overall speaker verification system is also improved by using histogram equalisation as a normalisation technique previous to fusing the features by SVM.

1 Introduction

One of the central issues addressed by automatic speaker recognition research is to find those features that convey speaker identity. Traditionally, automatic speaker recognition systems have relied mostly on low-level characteristics by using short-term features related to the spectrum of the voice, associated mainly to the physical traits of the vocal apparatus. However, humans rely on several linguistic levels contained in the speech signal in order to identify people from voice alone [1]: the voice timbre, a characteristic laugh, specific and repeatedly used words, etc. In

contrast to the spectral level, these linguistic features are mainly related to the learned habits and style.

Since this linguistic information plays an important role in the human recognition process, there is reason to believe, as recent studies have demonstrated [2-8], that it can add complementary knowledge to the traditional spectrum-based recognition systems, improving their accuracy. Moreover, they appear to be more robust to acoustic degradations from channel and noise effects [9, 10].

In the current paper, prosodic information is first added to a traditional spectral system in order to improve their performance, finding and selecting appropriated characteristics related to the human speech prosody, and combining them with the traditional spectral features. Such prosodic characteristics include parameters related to the fundamental frequency in order to capture the intonation contour, and other parameters such as the length of the speech segments to represent the speaker speech tempo.

Apart from the above-mentioned features, there may be many more characteristics that provide complementary information, being of great value for the speaker recognition task. The main objective of this paper is to use jitter and shimmer as additional acoustic features for speaker verification experiments. These features are related to the way how the speech is produced. Specifically, jitter and shimmer have been largely used to detect voice pathologies and to identify the age and gender of the speakers [11], which leads to think that they can be of usefulness in the speaker recognition task.

Prosodic and other additional features are especially useful when combined with the short-term spectral parameters [7, 8]. Therefore, the current paper focuses also on the combination of such features in order to improve the overall performance of the system. To this end, some existing normalisation and fusion methods are implemented in order to find those techniques with which the best performance is obtained. Moreover, a histogram equalisation of the scores is tested in order to improve the overall system performance.

This paper is structured as follows. Next, the spectral and prosodic parameters used in the baseline systems are described. In section 3, jitter and shimmer measurements are introduced. Section 4 deals with the normalisation and fusion techniques used in the current experiments. In section 5, the verification experiments and results are presented and discussed. Finally, conclusions of the current work are found in section 6.

2 Spectral and Prosodic Parameters

2.1 Voice spectrum-based parameters

Although the features extracted from the speech signal can be related to both source and filter processes, speaker recognition systems have tended to use only the filter features. These parameters, referred to as the spectral level of speech, relate to the physiology of the vocal tract and to the learnt articulatory configurations that shape the specific speech sounds [12-14].

The most commonly used parameters in the state-of-the-art speaker and speech recognition technologies are the Mel-frequency cepstral coefficients (MFCC) [15, 16]. However, the Frequency Filtering (FF) method [17] uses a first order filter $H(z) = z - z^{-1}$ instead of the Discrete Cosine Transform in the MFCC extraction process. This filter consists in a simple subtraction of the energies of two bands to compute each parameter, so that the computational cost is lower and the resulting parameters remain in the frequency domain. In most of the experiments performed in [18, 19], these parameters give comparable or even better results than Mel-cepstrum coefficients.

2.2 Prosody-based parameters

In order to recognise others with voice, humans use other speech sources like lexical terms, prosody or phonetics, which are related to the use of linguistic cues derived from language. Since speech carries this sort of information, some speaker recognition systems have also begun to use the source parameters together with the filter parameters. These source parameters relate

mainly to the fundamental frequency and power of the speech and, in turn, to the prosody of the spoken phrases [7, 8, 20].

Prosody is conveyed through three different elements: intonation, rhythm and stress [21]. It is well-known that prosody plays an important role in the speech act and communication in everyday speech [22, 23]; the speech, in turn, becomes adjustable to the the particularities of the speaker, so that each person may use distinct variations of tone, intensity and rhythm in their speech production. Prosodic variations apply normally to more than one phoneme: syllables, words, phrases, clauses, etc. Since phonemes are known as speech segments in linguistic terms, these prosodic elements are also know as 'suprasegmental features', and they are usually analysed over sequences of segments or entire syllables [21, 24].

The prosodic features used in this paper are inspired by the previous works of [25] and [7] and extracted using the manually corrected word-level transcriptions of the entire Switchboard-I corpus [26], and the Praat software for acoustic analysis [27]. These features, listed below, include features related to word and segmental duration and fundamental frequency. They are computed for each word and then averaged over all words.

Features related to word and segment duration

– Logarithm of number of frames per word.

– Fraction of voiced frames within each word.

Features related to fundamental frequency

F0-related features are estimated with Praat, by performing an acoustic periodicity detection based on a cross-correlation method using a Hanning window with a physical length of 40/3 ms and a shift of 10/3 ms:

– Logarithm of mean F0.

– Logarithm of maximum F0.

– Logarithm of minimum F0.

– Logarithm of the range of F0 (maximum F0 – minimum F0).

– F0 slope computed as: (last F0 – first F0)/(number of frames).

– Mean slope of the stylised F0 contour, using a 2 semitones frequency resolution [27].

Logarithmic transformations are performed on some features so that the distribution of values will look more Gaussian [28, 29]. Figure 1 illustrates this fact for the frames per word feature distribution. Values of skewness and kurtosis for all the transformed distributions yield within the Gaussianity limits, and they were more related to a normal distribution than the original distributions. In the example illustrated in Figure 1, skewness and kurtosis (which have 0 and 3 values in normal distributions) changed from 0.57 and 4.02 to 0.15 and 3.29, respectively.

3 Jitter and Shimmer

Fundamental frequency is determined physiologically by the number of cycles that the vocal folds do in a second. Jitter refers to the variability of F0, and it is affected mainly due to the lack of control of vocal fold vibration [30, 31]. On the other hand, vocal intensity is related with subglottic pressure of the air column, which, in turn, depends on other factors such as amplitude of vibration and tension of vocal folds [32]. Shimmer is affected mainly because of the reduction this tension and mass lesions in the vocal folds [30].

Both jitter and shimmer features have been largely used to detect voice pathologies (see, e.g. [33-35]). 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. More recently, they have also been used to determine the classification of human speaking styles [36] and the age and gender of the speakers [11]. Absolute jitter values, for instance, are found larger in males, while relative jitter values are larger in females [37]. Moreover, F0 and amplitude inestability increases with the aged voice, resulting in greater jitter and shimmer values and leading to tremor and increased hoarseness [38]. However, the ability of these features for age classification is normally reduced to two age intervals [39].

A study presented in [40] showed that jitter and shimmer originating from the glottal source are altered by the influence of the vocal tract, as stated in [35]. Also in [35], a theoretical

model for the measured jitter and shimmer interaction and the effects of the vocal tract is proposed and compared to the measured results, which show that significant amounts of jitter and shimmer are introduced in the signal by the vocal tract filtering, and that this alteration is nearly an order of magnitude depending on the fundamental frequency. Moreover, [35] demonstrates that large relative values of jitter and shimmer can affect the measurement accuracy.

In this paper, jitter and shimmer have been analysed in order to test their usefulness in speaker verification. These features are normally measured for sustained vowels in the frame voice pathology detection; however, since voiced consonants are useful in the speaker recognition task, jitter and shimmer are measured for all the voiced speech segments in the current experiments. Both features have been extracted by using the Praat voice analysis software, which reports different kinds of measurements for both jitter and shimmer features, listed below.

3.1 Jitter measurements

Jitter (absolute): 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 F0 period lengths and N is the number of extracted F0 periods, as shown in Figure 2.

Jitter (relative): average absolute difference between consecutive periods, divided by the average period:

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

Jitter (rap): Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbours, divided by the average period:

$$Jitter(rap) = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left|T_i - \left(\frac{\left(T_i + T_{i-1} + T_{i+1}\right)}{3}\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{3}$$

Jitter (ppq5): 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:

$$Jitter(ppq5) = \frac{\frac{1}{N-4}\sum_{i=3}^{N-2}\left|T_i - \left(\frac{\left(T_i + T_{i-2} + T_{i-1} + T_{i+1} + T_{i+2}\right)}{5}\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{4}$$

3.2 Shimmer measurements

Shimmer (absolute): 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:

$$Shimmer(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(A_{i+1}/A_i\right)\right| \tag{5}$$

where A_i are the extracted peak-to-peak amplitude data and N is the number of extracted fundamental frequency periods, as shown in Figure 3.

Shimmer (relative): average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude:

$$Shimmer(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - A_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{6}$$

Shimmer (apq3): three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its neighbours, divided by the average amplitude:

$$Shimmer(apq3) = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left|A_i - \left(\frac{A_i + A_{i-1} + A_{i+1}}{3}\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (7)$$

Shimmer (apq5): 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:

$$Shimmer(apq5) = \frac{\frac{1}{N-4}\sum_{i=3}^{N-2}\left|A_i - \left(\frac{A_i + A_{i-2} + A_{i-1} + A_{i+1} + A_{i+2}}{5}\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (8)$$

Shimmer (apq11): 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:

$$Shimmer(apq5) = \frac{\frac{1}{N-10}\sum_{i=6}^{N-5}\left|A_i - \frac{\sum_{k=i-5}^{i+5} A_k}{11}\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (9)$$

4 Normalisation and Fusion Techniques

Individual systems can be combined in order to obtain a better overall performance [41]. This combination can be performed at different levels: the feature extraction level, the match score level and the decision level. Fusion at the match score level is usually preferred by most of the systems; but prior to combining the scores of the matchers into a single score, a normalisation

process need to be performed in order to transform all the scores of the individual matchers into a common domain [42]. Thus, score combination is a two-step process: normalisation and fusion [43-46].

4.1 Normalisation methods

In the current paper, the conventional **z-score** (ZS) technique has been used to normalise the scores [42]. ZS sets the mean of the normalised scores to zero and their variance to one, according to:

$$s_{zs} = \frac{a - mean(A)}{std(A)} \tag{10}$$

being *mean(A)* and *std(A)* the statistical mean and the standard deviation of the set of scores *A*, repectively, and *a* the individual score to normalise.

On the other hand, **histogram equalisation** (HE) has been applied in this paper to the score distributions as a normalisation technique, performing a matching of the cumulative distribution function (CDF) of a reference score distribution and the CDF of the variable to be transformed, as recently done by the authors in [47]. HE is a non-linear transformation that converts a probability distribution to another, in order to match all the statistics of two probability distributions, reducing the mismatch of the statistics of two signals [48, 49]. This technique has also been developed for speech recognition adaptation approaches and correction of non-linear effects [50, 51], and it has also been applied to the acoustic features in order to improve the robustness of a speaker verification system by reducing the mismatch between training and test conditions and the additive noise and channel and transducer effects [49, 52].

4.2 Score-level fusion techniques

In the context of a verification task, two distinct approaches to score-level fusion can be considered: the combination approach and the classification approach [42]. The first one formulates the score fusion as a combination problem, where the individual matching scores are

combined using simple arithmetic or rule operations in order to generate a single scalar score, which is then used to make the final decision [53, 54]. In **matcher weighting** (MW) fusion, for instance, each unimodal score is weighted by a factor proportional to its recognition rate, so that the weights for more accurate matchers are higher than those of less accurate matchers [44]. When using EERs, for instance, the final score is expressed as:

$$u_{MW} = \sum_{i=1}^{N} w_i s_i, \qquad \text{where} \qquad w_i = \frac{1/EER_i}{\sum_{i=1}^{N} 1/EER_i} \tag{11}$$

being w_i and s_i the weighting factor and the individual score for the *i*-th modality, and N the number of modalities.

On the other hand, **support vector machines** (SVM) is one of the most currently used fusion techniques based on the classification approach, where the scores obtained by individual classifiers are seen as input patterns to be labelled as 'accepted' or 'rejected' [55, 56]. The SVM algorithm finds an optimal separating hyperplane (determined by the 'support vectors') that splits input data in two classes, maximising the distance of the hyperplane to the nearest data points of each class. Since data are normally not linearly separable, an extension to non-linear boundaries is achieved by using specific functions called kernel functions [57]. The kernel function used in the current experiments is a radial basis function expressed as:

$$K(x_i, x_j) = \exp\left[-\frac{1}{2} \left(\frac{\left\| x_i - x_j \right\|}{\sigma} \right)^2 \right]. \tag{12}$$

5 Recognition Experiments

First, this section focuses on the use of prosody in speaker recognition, and the combination with spectral information in order to improve the overall performance of a verification system. The baseline spectral system used in the current experiments is introduced in section 5.1. The performance of the individual prosodic parameters is shown in section 5.2. In section 5.3, these prosodic features are combined with the spectral parameters.

The current paper focuses also on the improvement of a prosodic and voice spectral verification system by introducing new features based on jitter and shimmer measurements. In section 5.4, both jitter and shimmer are introduced in more detail, and several methods to measure them are described. Both features are also used to perform some speaker recognition experiments again over the Swichboard-I database. In section 5.5, some selected jitter and shimmer measurements are used in combination with prosodic and short-term spectral parameters.

A conventional normalisation is applied before MW fusion in sections 5.3 and 5.4. In section 5.5, SVM are also used in the fusion process. Although SVM seem to be used alone without any prior normalisation; however, an intrinsec min-max normalisation is, in these cases, normally included. The experiments performed in this section using merely SVM fusion are thus previously min-max normalised, and when using HE as normalisation technique, a 100-interval equalisation will be applied.

5.1 Spectral baseline system

The spectrum-based recognition system used in this work is a 32-component GMM-UBM, with short-term feature vectors consisting of 20 Frequency Filtering parameters, including 20 corresponding delta and acceleration coefficients, a frame size of 30 ms and a shift of 10 ms.

All the verification experiments described in this paper have been performed over the Switchboard-I conversational speech database, which is a collection of 2430 two-sided conversations among 543 speakers from all areas of the United States, recorded in different telephone sessions. The whole speech conversations have a duration of three to ten minutes (so that each conversation side consists of three to four minutes on average), and the digital version of the speech signals was collected directly and automatically from the telephone network.

In the current paper, splits 1-3 of the database have been used as a training set for the speaker models, consisting of 135 registered speakers (67 males and 68 females). Each speaker model was trained with 8 conversation sides of the database. The UBM includes 116

conversation sides corresponding to 116 speakers taken from the complementary set of the database (splits 4-6), and gender balanced.

The system was tested using one conversation-side for each test trial, according to the NIST's 2001 Extended Data Task, with a total number of 1860 test trials (672 clients and 1188 impostors). In this task, the impostors were taken from the set of registered speakers, and some cross-sex trials were included. By using this experimental setup, the Equal Error Rate obtained in the spectrum-based speaker recognition system equals **10.1%.**

5.2 Prosodic system

A feature vector was obtained by using the eight characteristics described in 2.2. The system was tested using the *k*-nearest neighbour classifier (setting *k*=3) in the sum rule approach, comparing the distance of the test feature vector to the *k* closest vectors of the claimed speaker versus the distance of the test vector to the *k* closest vectors of the cohort speakers, and using two distance measures: the Euclidean distance and the symmetrised Kullback-Leibler divergence.

Table 1 shows the EER obtained for each prosodic feature using *k*=3. The best individual results are achieved in the features related to fundamental frequency, specially its mean values, and the Kullback-Leibler divergence as a distance measure.

Feature	Euclidean	Kullback-Leibler
Log (frames/word)	33.8	31.6
Voiced segments	33.5	30.1
Log (mean F0)	**24.8**	**20.4**
Log (max F0)	25.0	21.0
Log (min F0)	27.7	22.3
Log (range F0)	32.4	26.6
F0 slope	38.1	38.4
Stylised F0 slope	37.7	29.9

Table 1. EER (%) for each prosodic feature using two distance measurements and *k*=3.

5.3 Fusion of spectral and prosodic parameters

Next, the individual prosodic parameters are fused at the score level with the spectral-based system, using matcher weighting fusion combined with the conventional z-score normalisation. The spectral scores are obtained from the log likelihood function, while the prosodic scores are an inverse function of the distance measures, since these have inverse characteristics to that of log likelihood.

As in the previous experiments, splits 1-3 of the Switchboard-I database have been used to train the speaker models, and splits 4-6 have been set as cohort speakers. The complementary set consisting of splits 4-6 for training and splits 1-3 as cohort speakers is used as a developing set in order to obtain the weights for MW fusion and the statistical values needed for the normalisation phase.

The results after fusing the set of prosodic features with the spectral parameters are shown in Table 2, where the threshold of EER=10.1% corresponding to the performance of the spectral system is clearly outperformed.

Features	ZS-MW fusion
Prosodic	15.3
Prosodic + spectral	**7.7**

Table 2. EER (%) for prosodic features fused with spectral parameters, using z-score normalisation combined with MW fusion.

5.4 Jitter- and shimmer-based system

The recognition experiments described in this section have also been performed over the Switchboard-I database. A nine-feature vector was extracted for an acoustic system based on the nine jitter and shimmer measurements described in section 3. As in the F0-related features of the prosodic system, features were extracted using Praat, 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 experiment setup was the same used in the prosodic system, but taking only the Kullback-Leibler divergence as a distance measure, and using only z-score normalisation before MW.

First, in section 5.4.1, a verification task has been performed to test the performance of a system that uses solely jitter- and shimmer-based features. Additionally, in section 5.4.2, an identification task has been done in order to show how jitter and shimmer are useful to discriminate a voice of a particular speaker against peer speakers.

5.4.1 Verification experiments

Table 3 shows the EERs results for jitter and shimmer measurements, respectively, together with the combination of each measurement set. It seems 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 either.

In the case of shimmer measurements, their final fusion improves slightly the best isolated results (Shimmer (absolute)). Since all relative measurements of the same feature are highly correlated, only the relative measurement of shimmer giving the best EER is used: the Shimmer (apq3). To ensure that this measurement provides 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.

Jitter & shimmer measurement	EER (%)
Jitter (absolute)	**26.9**
Jitter (relative)	33.7
Jitter (rap)	34.2
Jitter (ppq5)	33.8
Jitter fusion	**29.2**
Shimmer (absolute)	**26.9**
Shimmer (relative)	28.9
Shimmer (apq3)	**28.1**
Shimmer (apq5)	32.9
Shimmer (apq11)	33.8
Shimmer fusion	**25.5**
3-JitShim	**22.5**

Table 3. EER (%) for jitter and shimmer measurements, isolated and combined using ZS-MW.

From now on, only three cycle-to-cycle variability measurements will be used as new features: the absolute measurement of jitter, the absolute measurement of shimmer, and one of the relative measurements of shimmer: the apq3. The EER of the combination of these set measurements, which will be referred to as the 3-JitShim system, equalled **22.5%.** The feature distributions of these three measurements are illustrated in Figure 4.

5.4.2 Identification experiments

Table 4 shows the results obtained after a closed set identification using the same 135 speakers utilised in the verification task, as well as two more identifications using closed sets of 100 and 50 speakers, all gender balanced. Since each speaker model has been trained with 8 conversation sides, the amount of remaining data for test is not large; therefore, in the 100 and 50 speakers closed sets, those speakers with less available test data have been skipped. The

three identification experiments have been performed with 704, 619 and 480 test trials, respectively.

Jitter & shimmer measurement	Identification Rate (%)		
	135 speakers	100 speakers	50 speakers
Jitter (absolute)	**6.4**	**8.2**	**21.6**
Jitter (relative)	6.3	8.2	20.4
Jitter (rap)	5.0	6.0	16.4
Jitter (ppq5)	4.0	5.5	14.8
Shimmer (absolute)	**8.5**	**9.9**	**27.0**
Shimmer (relative)	7.2	9.0	23.9
Shimmer (apq3)	7.0	8.9	24.3
Shimmer (apq5)	7.4	9.0	26.0
Shimmer (apq11)	3.8	5.3	18.6

Table 4. IR (%) for jitter and shimmer features considering different closed sets of speakers.

As in the verification results, it seems that the absolute measurements of jitter and shimmer are more useful to discriminate speakers than the relative ones. However, although the performance increases considerably when decreasing the dimension of the speakers set, these features are not good enough to be used on their own, and they need to be combined with other classical features.

5.5 Fusion of jitter and shimmer with prosodic and spectral features

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. Fusion of individual features is also performed at the score level, using the same experimental setup as in the previous sections.

First, all eight prosodic features used in the baseline system are combined with the three features of the 3-JitShim system, resulting in a new eleven-featured system. Second, the 3-JitShim system is added to the voice spectral baseline system. This allows to compare how complementary jitter and shimmer are to prosodic and spectral features, respectively. Finally, the 3-JitShim system is combined with both baselines, in order to see how the new features improve the speaker verification system. The results of these experiments are shown in Table 5 and their DET curves are plotted in Figure 5. The EERs obtained before using the 3-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.

Features	**ZS-MW fusion**	
	Baseline	**Baseline + 3-JitShim**
Prosodic	15.3	13.1
Spectral	10.1	8.6
Prosodic + spectral	7.7	6.8

Table 5. EER (%) for prosodic and spectral baseline systems before and after adding jitter and shimmer features using ZS-MW fusion.

The results and the DET curves plotted in Figure 5 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 3-JitShim system to the spectral system (15%), although when fusing 3-JitShim with the prosody based system the improvement is also considerable (14%). That suggests that the information provided by jitter and shimmer measurements to prosodic parameters and the information supplied to the spectral system are, in this case, equally complementary.

The speaker verification system based on prosodic and spectral parameters is also improved by adding the 3-JitShim system, as it can be seen in the DET curves plotted in Figure

6, achieving the lowest EER equalling 6.8%. In this case, the lower relative improvement with respect the ones achieved in prosodic and spectral baseline systems individually, may be do the use of MW technique, in which the weights associated to the best performing matchers are distributed and decreased at the expense of jitter and shimmer addition, or the existence of permanent external errors. In any case, jitter and shimmer features seem to be useful in speaker recognition and should be considered in future experiments.

Next, the 3-JitShim, the prosodic and spectral systems are combined using SVM fusion and HE as a normalisation technique. Table 6 shows the results obtained for each feature set and for the fusion of all feature sets. ZS-MS fusion is also shown for comparison, and average 90% Wald confidence intervals [58] are included in the last column.

Features	ZS-MW	SVM	HE-SVM	CI
3-JitShim	22.5	21.2	17.3	± 1.5
Prosodic	15.3	14.5	14.4	± 1.3
Spectral	10.1	10.1	10.1	± 1.1
All features	**6.8**	**6.7**	**6.1**	± 0.9

Table 6. EER (%) obtained in the fusion of jitter and shimmer, prosodic and spectral parameters using ZS-MW, SVM and HE-SVM.

Once again, SVM outperforms matcher weighting fusion. Moreover, the performance is, in all cases, even more improved when HE is used prior to SVM fusion as a normalisation technique, especially when applied over the 3-JitShim system.

6 Conclusions

Several works have demonstrated that the use of prosodic information helps to improve recognition systems based solely on spectral parameters. This fact has also been corroborated in sections 5.2 and 5.3 of the current paper, where a preliminary speaker verification system based

on prosodic features has been built in order to improve a voice spectrum-based verification system over the conversational Switchboard-I database. The experiments also show that Kullback-Leibler divergence used as a distance measure in the decision classifier outperformed, in almost all cases, Euclidean distance.

In sections 5.4 and 5.5, additional acoustic features, namely jitter and shimmer (which analyse the perturbation of fundamental frequency and waveform amplitude, respectively), have been used to improve a speaker verification system based on prosodic and spectral parameters. Both identification and verification results show that jitter and shimmer are not good enough to be used on their own, and that they need to be combined with other classical features. The current experiments have shown that jitter and shimmer can be used to provide complementary information to both spectral and prosodic systems, and that the absolute measurements of both jitter and shimmer parameters seem to be more speaker discriminative than their corresponding relative measurements.

The overall results vary depending on the fusion technique utilised in combining the involved parameters in the speaker recognition task. In this paper, the use of support vector machines outperforms the results obtained by matcher weighting technique. Future work may rely, for instance, on applying this fusion in a SoA system [59]. In addition, the experiments have shown that the overall results can be improved by applying a histogram equalisation as a normalisation technique.

Acknowledgments

This work has been supported by the Spanish Government under grant AP2003-3598. The authors would like to thank Pascual Ejarque and Andrey Temko for his help in the fusion techniques used in this work.

References

[1] Schmidt-Nielsen, A., and Crystal, T.H.: 'Speaker Verification by Human Listeners: Experiments Comparing Human and Machine Performance Using the NIST 1998 Speaker Evaluation Data', Digital Signal Processing, 2000, 10, pp. 249-266.

[2] Sönmez, M.K., Shriberg, E., Heck, L., and Weintraub, M.: 'Modeling dynamic prosodic variation for speaker verification'. Proc. ICSLP, Sydney, Australia, November 1998.

[3] Doddington, G.: 'Speaker recognition based on idiolectal differences between speakers'. Proc. Eurospeech, Aalborg, Denmark, September 2001.

[4] Andrews, W. Kohler, M.A., Campbell, J., Godfrey, J. and Hernández-Cordero, J.: 'Gender-dependent phonetic refraction for speaker recognition'. Proc. ICASSP, Orlando, Florida, May 2002.

[5] Bartkova, K., Le-Gac, D., Charlet, D., and Jouvet, D.: 'Prosodic Parameter for Speaker Identification'. Proc. ICSLP, Denver, Colorado, September 2002.

[6] Weber, F., Manganaro, L., Peskin, B., and Shriberg, E.: 'Using prosodic and lexical information for speaker identification'. Proc. ICASSP, Orlando, Florida, May 2002.

[7] Peskin, B. Navrátil, J., Abramson, J., Jones, D., Klusáček, D., Reynolds, D.A., and Xiang, B.: 'Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02'. Proc. ICASSP, Hong Kong, China, April 2003.

[8] Reynolds, D.A., Andrews, W., Campbell, J., Navrátil, J., Peskin, B., Adami, A., Jin, Q., Klusáček, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., and Xiang, B.: 'The SuperSID project: exploiting high-level information for high-accuracy speaker recognition'. Proc- ICASSP, Hong Kong, China, April 2003.

[9] Carey, M.J., Parris, E.S., Lloyd-Thomas, H., and Bennett, S.: 'Robust prosodic features for speaker identification'. Proc. ICSLP, Philadelphia, Pennsylvania, October 1996.

[10] Atal, B.S.: 'Automatic speaker recognition based on pitch contours', Journal of the Acoustical Society of America, 1972, 52, pp. 1687-1697.

[11] Wittig, F., and Müller, C.: 'Implicit Feedback for User-Adaptive Systems by Analyzing the User's Speech'. Proc. ABIS-03, Karlsruhe, Germany, 2003.

[12] Rabiner, L.R., and Juang, B.H.: 'Fundamentals of Speech Recognition'. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1993.

[13] Campbell, J.P.: 'Speaker recognition: A tutorial', IEEE, 1997, 85, pp. 1437-1462.

[14] Gish, H., and Schmidt, M.: 'Text-Independent Speaker Identification', IEEE Signal Processing Magazine, 1994, 11, pp. 18-32.

[15] Davis, S.B., and Mermelstein, P.: 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', IEEE Transactions on Acoustic, Speech and Signal Processing, 1980, 28, pp. 357-366.

[16] Oppenheim. A.V., and Schafer, R.W.: 'From Frequency to Quefrency: A History of the Cepstrum', IEEE Signal Processing Magazine, 2004.

[17] Nadeu, C., Hernando, J., and Gorricho, M.: 'On the decorrelation of filter bank energies in speech recognition'. Proc. Eurospeech, Madrid, Spain, September 1995.

[18] Hernando, J., and Nadeu, C.: CDHMM speaker recognition by means of frequency filtering of filter-bank energies'. Proc. Eurospeech, Rhodes, Greece, September 1997, pp. 2363-2366.

[19] Abad, A., Nadeu, C., Hernando, J., and Padrell, J.: 'Jacobian Adaptation based on the Frequency-Filtered Spectral Energies'. Proc. Eurospeech, Geneva, Switzerland, 2003.

[20] Adami, A.G.: 'Modeling prosodic differences for speaker recognition', Speech Communication, 2007, 49, pp. 277-291.

[21] Tuson, J. (dir.): 'Diccionari de lingüística', Barcelona: Vox, 2000.

[22] Nooteboom, S.: 'The Prosody of Speech: Melody and Rhythm', Hardcastle, W.J., and Laver, J. (Eds.): 'The Handbook of Phonetic Sciences', Oxford: Blackwell Publishers Ltd., 1997, pp. 641-673.

[23] Wennerstrom, A.: 'The Music of Everyday Speech. Prosody and Discourse Analysis', Oxford: Oxford University Press, 2001.

[24] Dellwo, V., Huckvale, M., and Ashby, M.: 'How is individuality expressed in voice? An introduction to speech production and description for speaker classification' in Müller, C. (Ed.): 'Speaker Classification', Berlin: Springer, 2007, vol. I, pp. 1-20.

[25] Shriberg, E., Stolcke, A., Hakkani-Tur, D., and Tur, G.: 'Prosody-based Automatic Segmentation of Speech into Sentences and Topics', Speech Communication, 2000, 32, pp. 127-154.

[26] Godfrey, J.J., Holliman, E.C., and McDaniel, J.: 'Switchboard: Telephone speech corpus for research and development'. Proc. ICASSP, Alburquerque, New Mexico, April 1990.

[27] Boersma, P., and Weenink, D.: 'Praat: Doing phonetics by computer', Website: http://www.praat.org, 1992.

[28] Limpert, E., Stahel, W.A., and Abbt, M.: 'Log-normal Distributions across the Sciences: Keys and Clues', BioScience, 2001, 51, pp. 341-352.

[29] Sönmez, M.K., Heck, L., Weintraub, M., and Shriberg, E.: 'A Lognormal Tied Mixture Model of Pitch for Prosody-based Speaker Recognition'. Proc. Eurospeech, Rhodes, Greece, September 1997.

[30] Behlau, M., Madazio, G., Feijó, D., and Pontes, P.: 'Avaliação da Voz', in 'Voz - O Livro do Especialista', Rio de Janeiro: Revinter, 2001, vol. I, chapter 3, pp. 86-180.

[31] Wertzner, H.F., Schreiber, S., and Amaro, L.: 'Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders', Revista Brasileira de Otorrinolaringologia, 2005, 71, pp. 582-588.

[32] Behlau, M., and Pontes, P.: 'Avaliação e Tratamento das Disfonias'. São Paulo: Lovise, 1995.

[33] Wagner, I.: 'A new jitter-algorithm to quantify hoarseness: an exploratory study', Forensic Linguistics, 1995, 2, pp. 18-27.

[34] Kreiman, J., and Gerrat, B.R.: 'Perception of aperiodicity in pathological voice', Journal of the Acoustical Society of America, 2005, 117, pp. 2201-2211.

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

[36] Li, X., Tao, J., Johnson, M.T., Soltis, J., Savage, A., Leong, K.M., Newman, J.D.: 'Stress and Emotion Classification using Jitter and Shimmer Features'. Proc. ICASSP, Honolulu, Hawaii, April 2007.

[37] Ludlow, C.L., Coulter, D.C., and Bassich, C.J.: 'Relationships between vocal jitter, age, sex, and smoking', The Journal of the Acoustical Society of America, 1982, 71, pp55-56.

[38] Linville, S.E.: 'The Aging Voice', The ASHA Leader, 2004, pp. 19-21.

[39] Sadeghi Naini, A., and Homayounpour, M.M.: 'Speaker age interval and sex identification based on Jitters, Shimmers and Mean MFCC using supervised and unsupervised discriminative classification methods'. Proc. ICSP Guilin, China, 2006.

[40] Kröger, B.: 'Zur Auswirkung der Glottis-Sprechtakt-Kopplung auf die Stimmreinheit', Sprache-Stimme-Gehör, 1991, 15, pp. 139-142.

[41] Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K., and Senior, A.W. 'Guide to Biometrics', New York: Springer, 2004.

[42] Jain, A., Nandakumar, K., and Ross, A.: 'Score Normalization in Multimodal Biometric Systems', Pattern Recognition, 2005.

[43] Fox, N.A., Gross, R., Chazal, P., Cohn, J.F., and Reilly, R.B.: 'Person identification using automatic integration of speech, lip and face experts'. Proc. ACM SIGMM 2003 Multimedia Biometrics Methods and Applications Workshop, Berkeley, CA, 2003.

[44] Indovina, M., Uludag, U., Snelik, R., Mink, A., and Jain, A.: 'Multimodal Biometric Authentication Methods: A COTS Approach'. Proc. Workshop on Multimodal User Authentication, Santa Barbara, CA, 2003.

[45] Lucey, S., and Chen, T.: 'Improved audio-visual speaker recognition via the use of a hybrid combination strategy'. Proc. AVBPA, Guildford, UK, 2003.

[46] Wang, Y., Wang, Y., and Tan, T.: 'Combining fingerprint and voiceprint biometrics for identity verification: an experimental comparison'. Proc. ICBA, Hong Kong, China, 2004.

[47] Farrús, M., Ejarque, P., Temko, A., and Hernando, J.: 'Histogram Equalization in SVM Multimodal Person Verification'. Proc. ICB, Seoul, Korea, 2007.

[48] de la Torre, Á., Peinado, A.M., Segura, J.C., Pérez-Córdoba, J.L., Benítez, M.C., and Rubio, A.J.: 'Histogram Equalization of Speech Representation for Robust Speech Recognition', IEEE Transactions on Speech and Audio Processing, 2005, 13, pp. 355-366.

[49] Skosan, M., and Mashao, D.: 'Modified Segmental Histogram Equalization for robust speaker verification', Pattern Recognition Letters, 2006, 27, pp. 479-486.

[50] Hilger, F., and Ney, H.: 'Quantile based histogram equalization for noise robust speech recognition'. Proc. Eurospeech, Aalborg, Denmark, 2001.

[51] Balchandran, R. and Mammone, R.: 'Non parametric estimation and correction of non linear distortion in speech systems'. Proc. ICASSP, Seattle, Washington, May 1998.

[52] Pelecanos, J., and Sridharan, S.: 'Feature warping for robust speaker verification'. Proc. t ODYSSEY-2001, Crete, Greece, 2001.

[53] Kitter, J., Hatef, M., Duin, R., and Matas, J.: 'On combining classifiers', IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20, pp. 226-239.

[54] Rodríguez-Liñares, L., García-Mateo, C., and Alba-Castro, J.L.: 'On combining classifiers for speaker authentication', Pattern Recognition, 2003, 36, pp. 347-359.

[55] Cristianini, N., and Shawe-Taylor, J.: 'An introduction to support vector machines (and other kernel-based learning methods)'. Cambridge University Press, 2000.

[56] Hearst, M.A.: 'Trends and Controversies: Support Vector Machines', IEEE Intelligent Systems, 1998, 13, pp. 18-28.

[57] Burges, C.J.C.: 'A tutorial on support vector machines for pattern recognition', Data Mining and Knowledge Discovery, 1998, 2, pp. 121-167.

[58] Newcombe, R.G.: 'Two-sided confidence intervals for the single proportion: Comparison of seven methods', Statistics in Medicine, 1998, 17, pp. 857-872.

[59] Zhang, X., Wong, H., and Cheung, W.: 'A Privacy-Aware Service-oriented Platform for Distributed Data Mining'. Proc. International Conference on E-Commerce Technology and the International Conference on Enterprise Computing, Palo Alto, California, 2006.

List of figure captions

Figure 1. Original feature distribution related to the number of frames per word (a), and resulting distribution after applying a logarithmic transformation (b).

Figure 2. Jitter measurement for *N*=4 F0 periods.

Figure 3. Shimmer measurements for *N*=4 F0 periods.

Figure 4. Feature distributions for the three measurements included in the 3-JitShim system.

Figure 5. DET curves for prosodic and spectral systems before and after adding jitter and shimmer features.

Figure 6. DET plot showing the improvement of the baseline system after adding jitter and shimmer.

Figures

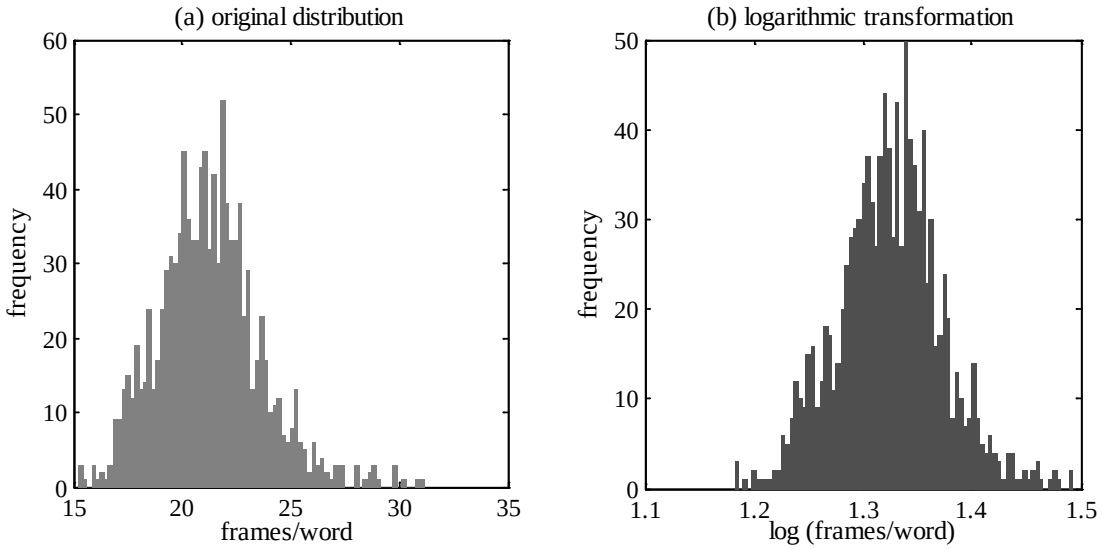

Figure 1. Original feature distribution related to the number of frames per word (a), and resulting distribution after applying a logarithmic transformation (b).

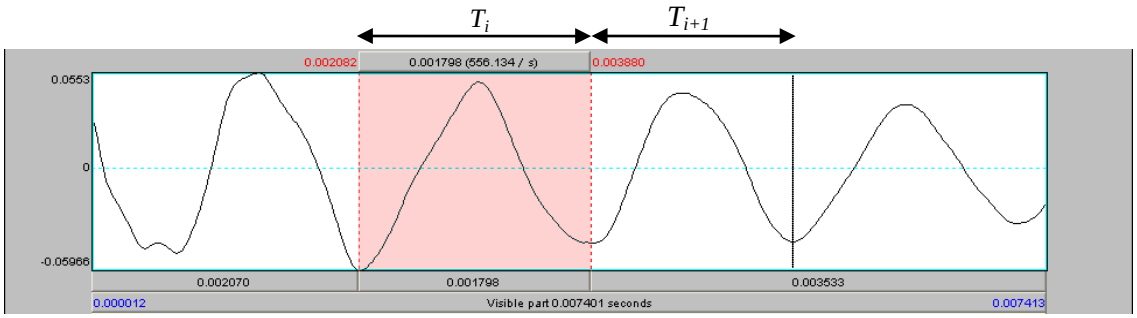

Figure 2. Jitter measurement for *N*=4 F0 periods.

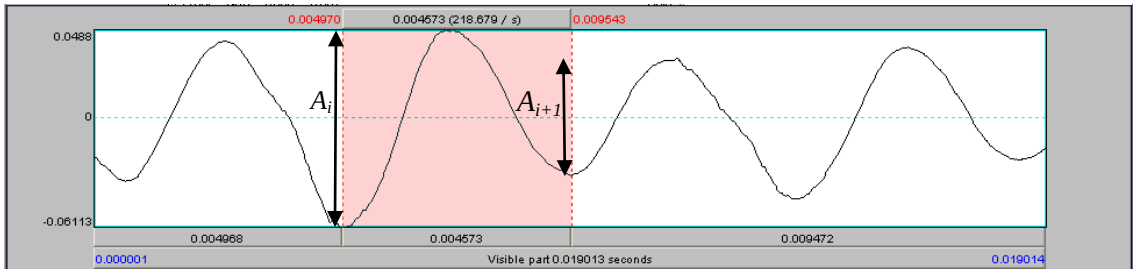

Figure 3. Shimmer measurement for *N*=4 F0 periods.

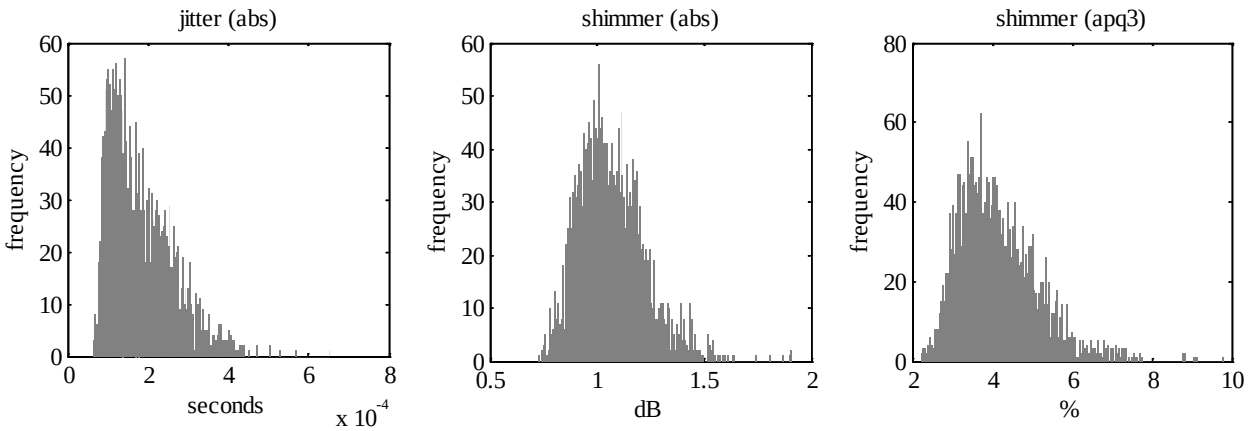

Figure 4. Feature distributions for the three measurements included in the 3-JitShim system.

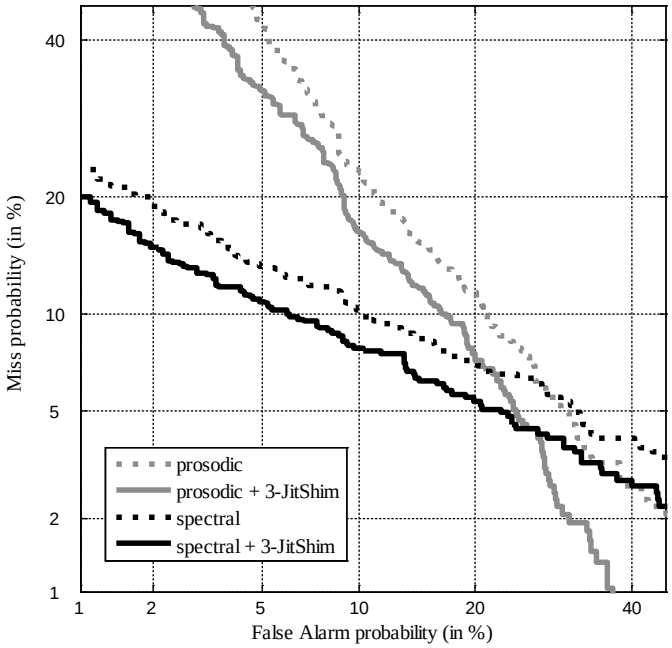

Figure 5. DET curves for prosodic and spectral systems before and after adding jitter and shimmer features.

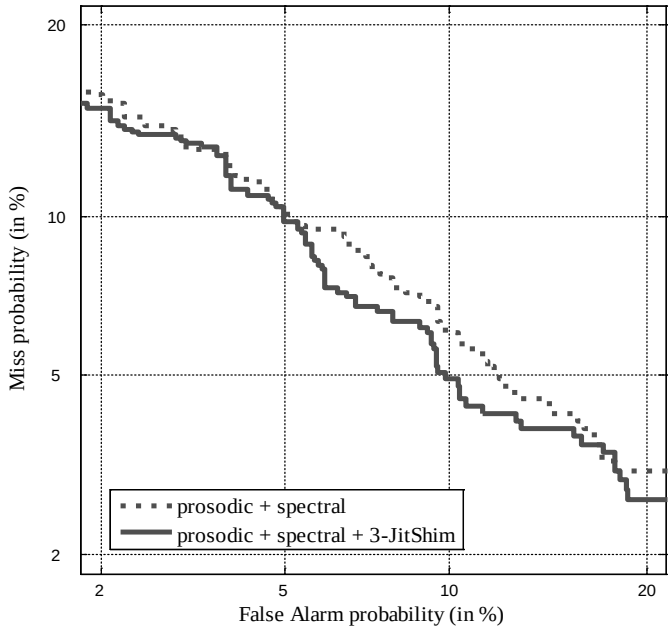

Figure 6. DET plot showing the improvement of the baseline system after adding jitter and shimmer.