Speaker Recognition with Spectral Dimension Features of Human Voices for Personal Authentication

Wen-Shiung Chen
VIPCCL, Dept. of Electrical Engineering, National Chi Nan University, Nan-Tou, Taiwan

Jr-Feng Huang
VIPCCL, Dept. of Electrical Engineering, National Chi Nan University, Nan-Tou, Taiwan

Abstract – Biometric recognition is more and more important due to security applications all over the world. Mobile phone becomes popular in recent years. Therefore, voice recognition on mobile devices for recognizing a speaker’s identity plays a potential role. This paper presents a speaker recognition method which combines a non-linear feature, named spectral dimension (SD), with Mel Frequency Cepstral Coefficients (MFCC). In order to improve the performance of the proposed scheme, the Mel-scale method is adopted for allocating sub-bands and the pattern matching is trained by Gaussian mixture model. Some problems related to spectral dimension are discussed and the comparison with other simple spectral features is made. We observe that our proposed methods can improve the performance in different components. For instance, speaker verification combining MFCC with our proposed SD features gives a good performance of EER=2.31% by 32_Multi-GMM. The relative improvement of about 22% may be achieved, which is better than the method that is based only on MFCC with EER=2.96%.

Index Terms – Biometric Recognition, Personal Authentication, Speaker Identification, Speaker Verification, Fractal Dimension, Spectral Dimension.

1. INTRODUCTION

Automatic speaker recognition (ASR) encompasses verification and identification [1]. Although the research of speech processing has been developed for many years, it still suffers from some problems, such as human and environmental factors. That ultimately limits ASR performance. Nevertheless, ASR is still the most natural and economical method for biometric authentication, and still needs more improvement.

Feature extraction is the kernel part in a biometric recognition system, which is the main concern of this paper. The speech features encompass high-level and low-level parts. The high-level features are related to dialect, speaker style and emotion state that are not always adopted due to difficult extraction [2]. The low-level features are related to spectrum, which is easy to be extracted, are always applied to ASR. The simplest feature is fundamental frequency [3]. There are some useful speech features, such as linear prediction coefficients residual signal (LPCRS), fractal dimension and other simple spectral features. Some researchers extract the features from LPCRS and wavelet transformation [4]. Fractal dimensions, such as box-counting dimension and Minkowski Bouligand dimension, can extract the properties of speech graph to complete speech and speaker recognition [5]-[7]. Correlation dimension and Lyapunov dimension can extract complexity from phase space, which is transformed in time domain signal with nonlinear dynamic methods [8]-[11]. A few researchers utilize GMM to gather statistical properties from phase space [12]. The fractal dimensions mentioned above have been applied to speech recognition, speaker recognition, biomedical signal processing and other signal processing. There are still unfamiliar fractal dimensions, such as spectral dimension and variance dimension, which have been applied to some related fields. The latter has even been applied to segment speech signals [13]. Then, the box counting dimension has been also applied to speech segmentation and enhancement [14]. The former has ever been applied to speech [10] and speaker recognition [11]. But the extraction method of spectral dimension is calculated by momentum theory, called critical exponent method (CEM). The fractal dimensions are always used in constrained text ASR systems. Hence, in the proposed system we extract the speech features by employing spectral dimension combined with traditional features MFCC to complete ASR task. The basic structure of the ASR system is illustrated in Figure 1.

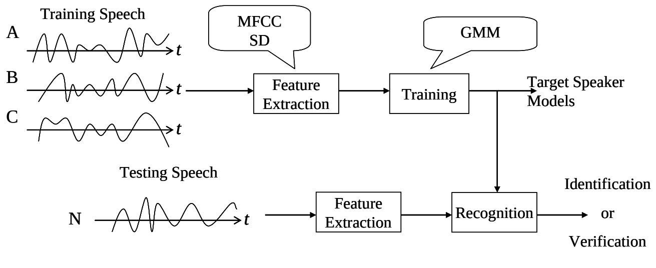

Figure 1 The basic structure of ASR system.

This paper is organized as follows. In Section 2, we introduce the definition of spectral dimension briefly and how to modify it in our ASR task. Then we also explain its rationality and propose our method. Section 3 gives the experimental results of the ASR system for identification and verification. Besides, we also compare with other simple spectral features. Section 4 finalizes this paper with some conclusions and future works.

2. PROPOSED METHOD

Spectral dimension is classified into transform fractal dimensions [15]. The time-domain signal can be transformed into its power spectrum density with spectral analysis techniques, such as the fast Fourier transform (FFT). If the power spectrum is broadband, with substantial power at low frequencies, it may originate from chaos. A signal $v(t)$ can be represented by either its energy spectrum or power spectrum. If we assume that the power spectrum density, $P(f)$, has the following power law form:

$$P(f) \propto \frac{1}{f^{\beta}}, \tag{1}$$

then we use the exponent β to define the spectral dimension as

$$D_{\beta} = D_E + \frac{3-\beta}{2} \tag{2}$$

where $D_E = 1$ is the embedding Euclidean for the time series. The β can be calculated by the following equations:

$$P(f) = k \cdot \frac{1}{f^{\beta}} \tag{3}$$

$$\begin{aligned} \log(P(f)) &= \log(k) + \log\left(f^{-\beta}\right) \\ &= -\beta \cdot \log(f) + \log(k) \end{aligned} \tag{4}$$

where k is a constant.

Simply, we may estimate the value of β by computing the slope of logarithm of spectrum vs. logarithm of frequency. It can be found with 1st-order polynomial fitting curve called least-squared method (LSM). Since the exponent β and spectral dimension are almost similar, this difference cannot affect the experimental performance. Then we take the exponent β into spectral dimension. Therefore, we adopt the exponent β to extract speech features. The colored noises, such as white, pink, brown and black, can also be represented by β.

According to our preliminary experiments, it could be found that the spectral dimension for representing speech property is not representative if it is computed directly from original entire spectrum. In order to improve the original SD, we adopted Mel-scale method, which is just a transformation of frequency based on human auditory perception, to segment the speech waveform in frequency domain [18]. Segmenting several overlapped sub-bands can retain consecutive property. We select twelve segmented sub-bands due to the number of MFCC. In this way, we do not consider the problem of weights. This segmentation technique for extracting spectral dimensions represents the property of entire spectrum perfectly, as shown in Figure 2. We name it *Mel-scale SD*. Accordingly we use the twelve spectral dimensions combining with twelve MFCC to form the new features. Consequently, the Mel-scale SD should gain some improvement on recognition performance.

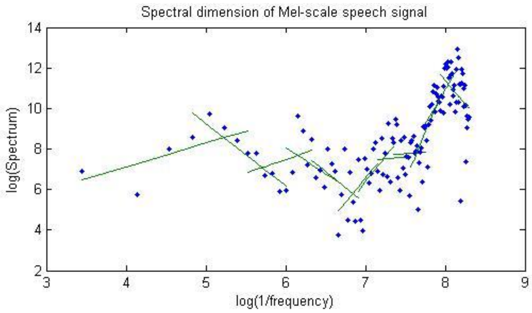

Figure 2 Mel-scale spectral dimension.

In general, MFCC and SD are embedded into the same feature vector, and GMM [1] is usually used in pattern recognition. The performance may not be better than MFCC due to improper normalization. If we do not use proper normalization, the performance of combining feature sets will be limited. Hence, we adopt multi-GMM to train individual features and deal with the above problem. Here we do not consider the problem of weighting, and assume that these features have potential discrimination. Each different feature set trains its own model separately, as shown in Figure 3. In this way, the normalization problem may be avoided.

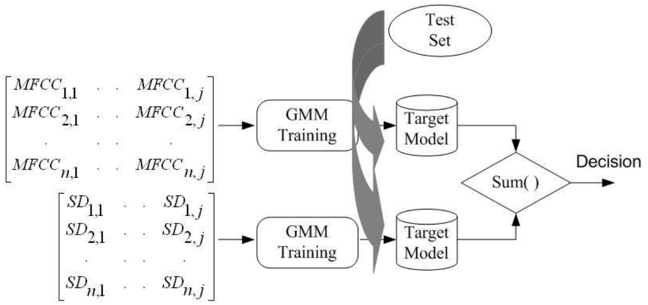

Figure 3 The Multi-GMM structure, where *n* is the dimension of the features and *j* is the number of frames.

We also discuss some issues about SD extraction. First, some researchers thought that speaker recognition does not need to extract features in complete spectrum. Then the linear prediction of spectral envelope, which is a smoothing procedure, was always adopted. This paper also extracted SD based on this approach and compared with our proposed SD. In order to discriminate easily, we named it *linear prediction spectral dimension* (LPSD). Second, some people consider that the SD extraction is only the slope of spectrum and it is easy to be calculated. It is still unknown about why we found SD between logarithm of spectrum and logarithm of frequency rather than spectrum and frequency. For logarithm of spectrum, it avoids large change in spectrum and is convenient to conceal channel noise. For logarithm of frequency, it can obtain *vocal tract length normalization* (VTLN) effect. In the literature, Sinha and Umesh [16] discussed VTLN in speech recognition. Here, we also provide a simple explanation for VTLN.

It is commonly assumed that the spectra of the same sound spoken by any two speakers are linearly scaled version of one and another due to differences in the vocal tract length. There are two simple models for VTLN, such as linear and non-linear scaling models. Especially, the non-linear scaling model is more suitable and represented as:

$$s_a(v) = S_A(f = h(v)) = s_b(v + \tau_{AB}) \qquad (5)$$

where τ_{AB} is a fixed translation factor. For the convenience of experimenting, the inverse-warping function adopted λ=log(v) and the fixed translation factor was neglected. Intuitively we could discover that the variation of frequency-axis range is quite small, but the variation of spectral amplitude-axis range is too large to perform well. Therefore, the original SD concept could be reasonable. So we also compare with the non-VTLN SD that directly captures the spectral variation in original frequency domain.

3. EXPERIMENTAL RESULTS

3.1. Experiments

Our ASR is performed and tested on AURORA2.0 that was published by European Telecommunications Standards Institute (ETSI) and was used to evaluate robust digital recognition. In this paper we only select clean sets from the AURORA2.0 database that contains clean digital series of 52 males and 57 females. Each speaker contains 77 sentences, and the contents are clean free-length English digital series, sampled at 8 KHz, with a resolution of 16 bits per sample. We apply these speech samples in our system. There are 39 sentences for training and 38 sentences for testing. The pre-processing only adopted end point detection due to the clean speech. Due to the speech production process and empirical experience, we pre-emphasize the speech before feature extraction. The length of the frame and overlap are 32 ms (256 samples) and 10.6 ms (85 samples), respectively, and a Hamming window is applied. 12-MFCCs without energy, first and second derivatives are extracted from speech. The acronyms in the following tables and figures are also stated here. The original spectral dimension is called as 1SD as extracted from whole frequency domain. The Mel-scale SD is called as SD. There are still some simple spectral features applied to speaker identification, such as spectral centroid (SC), spectral bandwidth (SBW), spectral band energy (SBE), spectral flatness measure, spectral crest factor (SCF), Renyi entropy (RE) and Shannon entropy [17]. They claimed that only part of spectral features could improve performance of the system. So we only compare the ones with the better features and another traditional feature, linear prediction cepstral coefficients (LPCC) with our proposed method.

3.2. Results of Speaker Verification

In the following we adopt Gaussian mixture model with universal background model (GMM-UBM) [18] to carry out the experiments of verification and each class owns one common UBM. It can be completed on lower components than identification. Each class has 38 real targets and 38×108=4,104 imposters. The performance of this task is evaluated by equal error rate (EER) that is determined when false acceptance rate (FAR) and false rejection rate (FRR) are equal. First, we will show the results that we discussed before, such as SD, multi-GMM, LPSD, and non-VTLN SD.

We discover that MFCC combining with 1SD cannot improve efficiently. Hence, we may conclude that it cannot represent the property of speech well. After adopting SD, the performances of other components are also better than baseline with embedding SD and MFCC into one vector, called non-multi GMM. In order to obtain further improvement, we adopt multi-GMM mentioned above in our work. The stable and better performance is discovered by multi-GMM. Thus, this pattern matching method will be adopted in verification. We present how essential LPSD is. However, the experimental results show that the performances are similar to our proposed SD. In addition to GMM_32, most works on different components have worse performance than before. Hence, the linear prediction is not essential. When we adopt non-VTLN SD, there are a few improvements in most components in addition to GMM_32. Compared to our proposed SD, these still do not perform well. Thus, the VTLN is of importance for our task. The above results are shown in Table 1 and Figure 4.

	GMM_8	GMM_16	GMM_32
MFCC	8.30%	5.15%	2.96%
MFCC+1SD (non-Mel scale)	8.21%	5.07%	3.02%
SD (Mel scale)	9.78%	6.23%	4.33%
MFCC+SD (non Multi-GMM)	7.84%	4.54%	2.70%

MFCC+SD (Multi-GMM)	6.70%	3.80%	2.31%
MFCC+LPSD	6.90%	3.83%	2.28%
MFCC+ Non-VTLN SD	7.45%	4.88%	2.98%

Table 1 Results of speaker verification

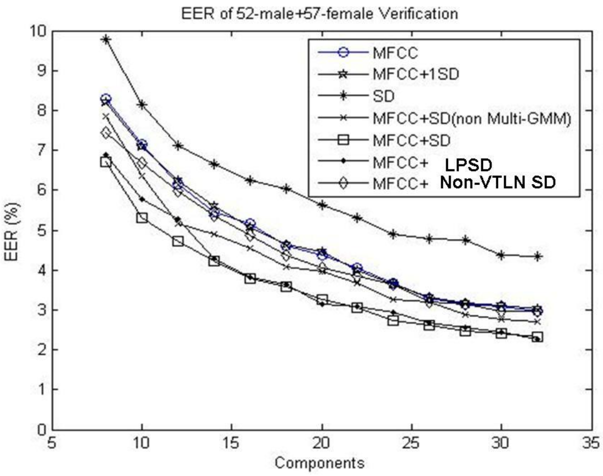

Figure 4 Speaker verification on different components.

As a result of the different length of testing utterances, the above results are calculated by an average EER. Hence, we could not describe the detection error tradeoff (DET) curves for all classes in detail. In order to describe a DET curve which can represent the score distribution of all classes in common threshold range, we have to calculate a new EER under this situation. The EERs occur at cross points on dash line. Figure 5 shows the DET curves of MFCC, MFCC combining with Mel scale SD, and Mel scale SD in GMM_32. Our methods could obtain improvement of EER performance, although not apparently.

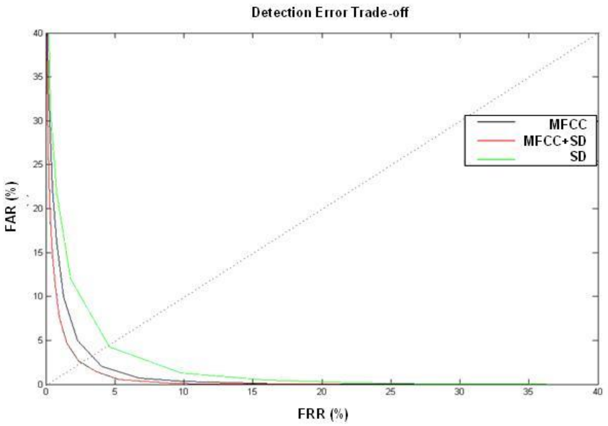

Figure 5 The DET curves for speaker verification.

We compare some simple spectral features that we mentioned earlier. There are a few spectral feathers that are beyond our anticipation, such as SBE and RE. Our SD method outperforms most simple spectral features in addition to SC. Although it is not the best, our method still has better improvement than the traditional LPCC. Besides, combing with SC can obtain a few improvements mostly. The comparison results are shown in Table 2 and Figure 6.

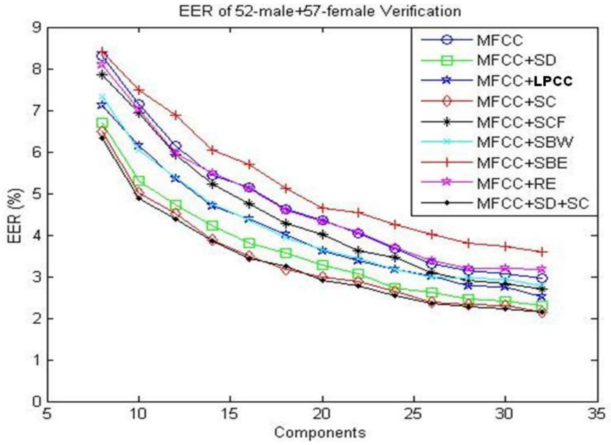

Figure 6 Speaker verification compared with other features on different components.

Algorithms	GMM_8	GMM_16	GMM_32
MFCC	8.30%	5.15%	2.96%
MFCC+SD	6.70%	3.80%	2.31%
MFCC+LPCC	7.11%	4.38%	2.50%
MFCC+SC	6.47%	3.49%	2.16%
MFCC+SCF	7.85%	4.75%	2.69%
MFCC+SBW	7.32%	4.35%	2.78%
MFCC+SBE	8.42%	5.69%	3.59%
MFCC+RE	8.09%	5.13%	3.16%
MFCC+SC+SD	6.33%	3.44%	2.16%

Table 2 Speaker verification compared with other features

3.3. Results of Speaker Identification

Identification belongs to close-set. Suppose that all testing speech signals belong to the known speakers. Since identification is more difficult than verification, it needs more components to achieve better performance. Hence, it is always completed on higher components. Then the performance of this task is evaluated in accuracy rate which is determined by how many test sets are correct.

The results show that MFCC combining with 1SD cannot improve efficiently. The reason is the same as before in this task. After adopting SD, the performances of other components are also improved more than baseline. In order to improve more efficiently, we also adopt multi-GMM. Its advantage is also discovered from the result. Therefore, the pattern matching method will be also adopted in following identification task. However, the experimental result of LPSD shows that the performances are similar to our proposed SD, but most works on different components are still worse than the proposed SD. Hence, the linear prediction method is not essential and may ignore some information related to speakers. When we adopt non-VTLN SD, the accuracy gets worse. In this task, we could make a conclusion that the VTLN used in identification is more important than in verification. The above results are shown in Table 3 and Figure 7.

Algorithms	GMM_32	GMM_64	GMM_128
MFCC	89.43%	92.30%	93.26%
MFCC+1SD (non-Mel scale)	89.70%	91.89%	93.14%
SD (Mel scale)	82.93%	85.01%	85.13%
MFCC+SD (non Multi-GMM)	90.99%	93.31%	93.70%
MFCC+SD (Multi-GMM)	91.77%	93.55%	94.50%
MFCC+LPSD	91.45%	93.46%	94.40%
MFCC+ Non-VTLN SD	89.14%	91.60%	92.90%

Table 3 Results of speaker identification

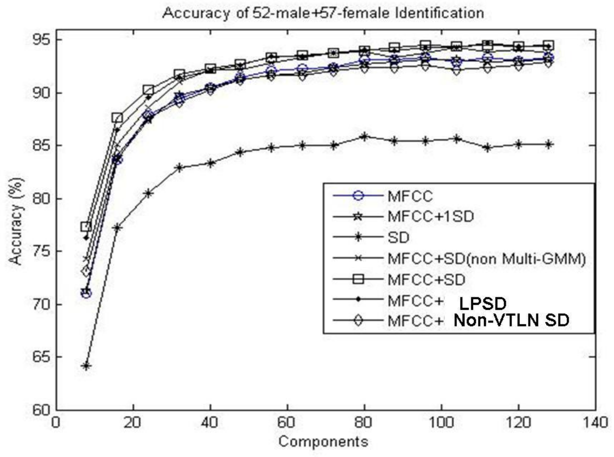

Figure 7 Speaker identification on different components.

The comparison results reveal that SBE, RE and SCF perform not well. We found that the SCF cannot perform well here. Our proposed SD still outperforms most simple spectral features in addition to SC. Although it is not also the best one, our proposed SD still has better improvement than the traditional LPCC. Besides, combing with SC can obtain a few improvements mostly. The performance of identification is more significant than verification. The comparison results are shown in Table 4 and Figure 8.

	GMM_32	GMM_64	GMM_128
MFCC	89.43%	92.30%	93.26%
MFCC+SD	91.77%	93.55%	94.50%
MFCC+LPCC	91.04%	93.17%	94.13%
MFCC+SC	92.06%	93.72%	94.54%
MFCC+SCF	89.72%	91.77%	92.93%
MFCC+SBW	89.93%	92.47%	93.34%
MFCC+SBE	86.99%	90.32%	91.65%
MFCC+RE	88.53%	91.12%	92.06%
MFCC+SD+SC	92.64%	94.13%	94.88%

Table 4 Speaker identification compared with other features

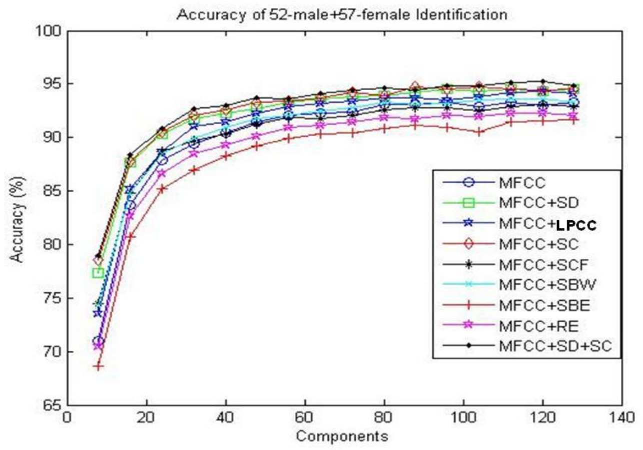

Figure 8 Speaker identification compared with other features on different components.

3.4. Discussion

By comparing different tasks, we could observe that there are almost the same issues in our results. The SC related to the local shape of spectral could have the best performance, since acoustic structure contains more useful information than acoustic variation. The results of combining with SCF related to local spectral formant do not gain a significant improvement, and even get worse in speaker identification. It seems that these formants may produce confusing result, which is investigated in [3]. The SBE is not suitable to our speaker recognition tasks due to the volume of voice in continuous digital utterance. However the SBW related to SC and SBW could get a little improvement. Since the RE is suitable for detecting voiced and unvoiced components of speech originally, it does not perform well mostly in continuous digital utterance. The fusion of shape and variation could provide complementary effect further.

4. CONCLUSION

This paper proposed a personal authentication for speaker recognition based on Mel-scale spectral dimension and MFCC. Assuredly, the proposed methods could gain more improvements. The rationality of spectral dimension has also been mentioned here, and we adopt non-VTLN SD to prove it. In pattern recognition, the advantage of multi-GMM is discovered. Comparing with other spectral features could have better performance in addition to SC. Fusion of MFCC, SD and SC could improve a little recognition rate further.

In this paper, we did not take into account the problem of weightings. We will compute the contribution of different speech features. Besides, our ASR only performs on clean digital data. We assume that the partial spectral variation may stable in noise environment. Therefore, we will also evaluate it on additive noise data and on different type of data. Then how to find optimal components related to speakers of GMM should be completed in the future. Besides, we will search some knowledge-based features to replace partial data-driven approaches.

REFERENCE

[1] J. P. Campbell Jr., “Speaker recognition: A tutorial,” *Proc. of the IEEE*, vol. 85, no. 9, 1997, pp. 1437-1462.

[2] J. P. Campbell, D. A. Reynolds and R. B. Dunn, “Fusing high- and low-level features for speaker recognition,” in *Proceedings of Eurospeech*, Geneva, 2003, pp. 2665-2668.

[3] V. Pitsikalis and P. Maragos, “Nonlinear speech processing applied to speaker recognition,” in *European Cooperation in the Field of Scientific and Technical Research*, 2002.

[4] N. Zheng, T. Lee and P. C. Ching, “Integration of complementary acoustic features for speaker recognition,” *IEEE Signal Processing Letters*, vol. 14, no. 3, pp. 181-184, 2007.

[5] P. Maragos and A. Potamianos, “Fractal dimensions of speech sounds: computation and application to automatic speech recognition,” *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1925-1932, Mar. 1999.

[6] M. Yao, J. Hu and Q. Gu, “A mixed parameter method based on MFCC and fractal dimension for speech recognition,” in *IEEE Int. Conf. on Information Acquisition*, 2006, pp. 20-23.

[7] F. V. Nelwamondo, U. Mahola and T. Marwala, “Improving speaker identification rate using fractals,” in *International Joint Conference on Neural Networks (IJCNN)*, 2006, pp. 3231-3236.

[8] V. Pitsikali and P. Maragos, “Speech analysis and feature extraction using chaotic models,” in *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2002, vol. 1, pp. 533-536.

[9] V. Pitsikalis and P. Maragos, “Filtered dynamics and fractal dimensions for noisy speech recognition,” *IEEE Signal Processing Letters*, vol. 13, no. 11, Nov. 2006, pp. 711-714.

[10] H. Koga and M. Nakagawa, “Chaotic and fractal properties of vocal sounds,” *Journal of the Korean Physical Society*, vol. 40, no. 6, pp. 1027-1031, Jun. 2002.

[11] P. D. Augusto and C. Barone, “Speaker identification using nonlinear dynamical features,” *Chaos, Solitons, and Fractals*, vol. 13, pp. 221-231, 2002.

[12] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, F. M. Roberts and J. Ye, “Statistical models of reconstructed phase spaces for signal classification,” *IEEE Trans. on Signal Processing*, vol. 54, no. 6, pp. 2178-2186, Jun. 2006.

[13] W. Grieder and W. Kimner, “Speech segmentation by variance fractal dimension,” in *Canadian Conference on Electrical and Computer Eng.*, 1994, vol. 2, pp. 481-485.

[14] L. Chen and X. Zhang, “New methods of speech segmentation and enhancement based on fractal dimension,” in *5th IEEE International Conference on Signal Processing (WCCC-ICSP)*, 2000, vol. 1, pp. 281-284.

[15] W. Kinsner, “A unified approach to fractal dimensions,” in *4th IEEE International Conference on Cognitive Informatics (ICCI2005)*, U. of California, Irvine, USA, 2005, pp. 58-72.

[16] R. Sinha and S. Umesh, “A shift-based approach to speaker normalization using non-linear frequency-scaling model,” in *Proc. of the International Speech Communication Association*, 2008, vol. 40, pp. 191-202.

[17] D. Hosseinzadeh and S. Krishnan, “On the use of complementary spectral features for speaker recognition,” *EURASIP J. on Advances in Signal Processing*, 10 pages, 2008.

[18] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. O. Garcia, D. Petrvovska-Delacretaz and D. A. Reynolds, “A tutorial on text-independent speaker verification,” *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430-451, 2004.