Automatic speaker verification based on fractional Brownian motion process

R. Sant'Ana, R. Coelho and A. Alcaim

A novel text-independent verification system based on the fractional Brownian motion (*M_dim_fBm*) for automatic speaker recognition is presented. The performance results of the *M_dim_fBm* were compared to those achieved with the Gaussian mixture models (GMM) classifier using the mel-cepstral coefficients. A speech database, obtained from fixed and cellular phones, uttered by 75 different speakers was used. The results have shown the superior performance of the *M_dim_fBm* classifier in terms of recognition accuracy. In addition, the proposed verification scheme employs a much simpler modelling structure as compared to the GMM.

Introduction: In this Letter, we propose a text-independent speaker verification system which incorporates a new classification scheme based on the fractional Brownian motion (fBm) stochastic process. The fBm [1] is a mono-fractal stochastic process, i.e. it uses a single value of the Hurst (H) parameter [2]. The H parameter ($0<H<1$) represents the samples' time-dependence or scaling degree of a stochastic process.

According to the value of H and so the decaying rate of the auto-correlation coefficient function $\rho(k)$ ($-1<\rho(k)<1$) as $k\rightarrow\infty$, a stochastic process shows the presence of: (i) anti-persistence – $0<H<1/2$ where the auto-correlation function rapidly tends to zero and $\sum_{k=-\infty}^{\infty}\rho(k)=0$; (ii) short-range dependence (SRD) – $H=1/2$ where the auto-correlation function $\rho(k)$ exhibits an exponential decay to zero, such that $\sum_{k=-\infty}^{\infty}\rho(k)=c$, where $c>0$ is a finite constant and (iii) long-range dependence (LRD) – $1/2<H<1$ where the auto-correlation function $\rho(k)$ is a slowly-vanishing function which means a strong time-dependence even between samples that are far apart. In this case, we have $\sum_{k=-\infty}^{\infty}\rho(k)=\infty$.

To be suitable for applications in ASR systems we developed a new classification scheme called multi-dimensional fractional Brownian motion (*M_dim_fBm*). The proposed classifier is obtained from the set of H parameters, means and variances computed from any speech feature matrix. The *M_dim_fBm* classifier models the speech signal features considering their time-dependence or scaling characteristics. We have compared the performance of the *M_dim_fBm* to those achieved with the GMM [3] classifier using the mel-cepestral coefficients.

For fractal or self-similar processes only, we can relate the H parameter to a fractal dimension (D_h) [1] through the equation $D_h=2-H$. The fractal dimension was previously used in pattern recognition studies in [4] and [5]. In [6], the fractal dimension was applied for discriminating fricative sounds. A speaker identification system using cepstral coefficients is compared in [7] to a system based on the joint use of cepstral coefficients and the fractal dimension. These studies share the hypothesis that speech is a fractal signal. In this Letter, however, although we estimate the H parameter from the speech feature matrix, we do not assume that the speech signal is a fractal or self-similar signal.

Description of M_dim_fBm classifier: The *M_dim_fBm* model of a given speaker is generated according to the following steps:

1. Pre-processing: the feature matrix is formed from the input speech features. It contains c rows, where c is the number of feature coefficients per frame and l columns, where l is the number of frames.
2. Decomposition: for each row of the feature matrix we apply the wavelet decomposition and obtain the detail sequences where j is the decomposition scale and k is the coefficient index of each scale.
3. Parameters extraction/estimation: from each set of detail sequences obtained from each row of step 2, we estimate the mean, the variance and the H parameter. For the H parameter estimation we use the Abry-Veitch wavelet-based estimator proposed in [8].
4. Generation of fBm processes: using the random midpoint displacement (RMD) algorithm [1] and the three parameters computed in step 3, we generate the fBm processes. Therefore, we obtain c fBm processes.
5. Determining histogram and generating speaker model: we compute the histogram of each fBm process. The set of all histograms defines a c-dimensional fBm process which defines the speaker *M_dim_fBm* model.

In the phase of tests we use the histograms of the speaker *M_dim_fBm* model to compute the probability that a certain c-dimensional feature vector x belongs to that speaker. This is performed to the l feature vectors, resulting in l probability measures: $p_1, p_2, \ldots, p_l$. Adding these values, we obtain a measure of the likelihood that the set of feature vectors under analysis belongs to that speaker.

Experimental results: In this Section, we compare the results of the verification performance of the proposed *M_dim_fBm* system to those of the Gaussian mixture models (GMM) classifier. The database (BaseIME) used in our experiment is composed of 75 speakers (male and female). In fact, we have two databases: in one of them the speech signal was recorded from a fixed telephony channel and in the other one speech was obtained from a cellular telephony channel. Tests were applied to 20, 10 and 5 s speech segments. The best (upper limit) GMM performance is generally achieved for 32 Gaussians [3]. In our experiments, we have used 15 mel-cepstral coefficients for both classifiers and 32 Gaussians for the GMM classifier. Note that the feature matrix has $c=15$ rows, hence, we have a *M_dim_fBm* dimension equal to 15. A separate speech segment of 1 min duration was used to train a speaker model.

From several experiments, we have found that a good configuration for the H parameter estimation is given by the following specifications: (i) frame duration: 80 ms; (ii) Daubechies wavelets [9] with 12 coefficients; (iii) number of decomposition scales: 6; (iv) scaling region from 3 to 5.

The performance results for the text-independent speaker verification systems were obtained by varying the threshold and computing the miss (false rejection) and the false alarm (false acceptance) probabilities. These error probabilities are plotted using the detection error trade-off (DET) curves [10]. We have used as background the universal background model (UBM) model [11]. This one was constructed from speech material of 20 speakers that do not belong to the set of 75 speakers used for the testing experiments.

Figs. 1 and 2 show the DET curves for the *M_dim_fBm* and GMM based on 15 mel-cepstral coefficients for the speech database obtained from fixed phones, respectively. The results presented in these Figures show that the *M_dim_fBm* classifier in general presented better performance when compared to the GMM classifier. Note that the performance gains are substantial for a wide range of medium to low false alarm probabilities. It is important to remark that in most applications high false alarm probabilities must be avoided.

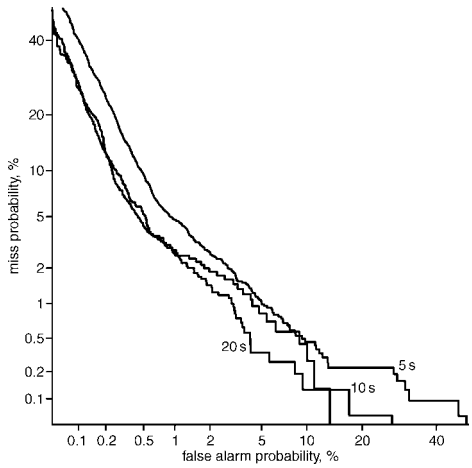

Fig. 1 *DET curves for systems based on mel-cepstral coefficients using M_dim_fBm classifiers for fixed phone*

Table 1 presents the equal recognition rates (ERR) for the operating point of the DET curve where $f_r=f_a$. This measure is given by $ERR=(1-EER)100\%$ where EER is the equal error rate usually employed in the literature. As we note the ERR is comparable for both systems. However, the DET curves show that for most of the operating points (miss probability × false alarm probability) the proposed classifier provides better results.

These results, along with the DET curves, corroborate the superior modelling procedure of the *M_dim_fBm* strategy for the speaker verification task. Moreover, the *M_dim_fBm* results were achieved

for a simpler model with dimension equal to 15. Each fBm is characterised by only three scalar parameters (mean, variance and H). Conversely, the GMM used 32 Gaussians, each one characterised by 1 scalar parameter, one mean vector and one covariance matrix, to achieve the performance results presented in Figs. 1 and 2. This means that the *M_dim_fBm* classifier yields a better modelling accuracy with a lower computational load.

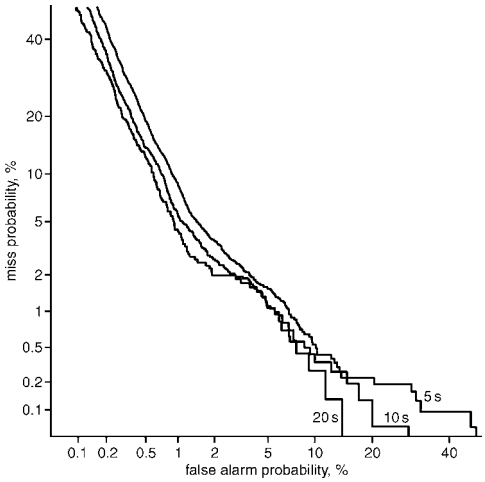

Fig. 2 *DET curves for systems based on mel-cepstral coefficients using GMM classifiers for fixed phone*

Table 1: ERR (%) of verification systems based on mel-cepstral coefficients, for speech signals recorded from fixed telephony and cellular telephony channels

Test duration	*M_dim_fBm* (fixed)	GMM (fixed)	*M_dim_fBm* (cel)	GMM (cel)
20 s	98.08	98.00	95.06	94.93
10 s	98.31	97.59	94.67	94.63
5 s	97.62	97.27	94.34	94.32

Conclusions: In this Letter we have presented a new classifier for text-independent speaker verification, the *M_dim_fBm*. The proposed classification approach is generated from the fractional Brownian motion stochastic process. We have shown that, as compared to the GMM classifier, the *M_dim_fBm* yielded the best overall recognition accuracy for the verification task. The results presented in this Letter show that the *M_dim_fBm* provides a more accurate and much simpler modelling strategy as compared to the GMM. We conclude, therefore, that the *M_dim_fBm* is a very attractive tool in the area of automatic speaker recognition systems and represents an important contribution due to its performance and simplicity.

20 May 2004
Electronics Letters online no: 20045090
doi: 10.1049/el:20045090

R. Sant'Ana and R. Coelho (*Electrical Engineering Department, Instituto Militar de Engenharia (IME), Praça General Tibúrcio 80, Praia Vermelha, Rio de Janeiro, Brazil*)

A. Alcaim (*CETUC/PUC-Rio, Rua Marquês de S. Vicente 225, Gávea, Rio de Janeiro, Brazil*)

References

1 Barnsley, M., *et al.*: 'The science of fractal images' (Springer-Verlag, New York Inc., USA, 1988)
2 Hurst, E.: 'Long-term storage capacity of reservoirs', *Trans. Am. Soc. Civil Engineers*, July 1951
3 Reynolds, D., and Rose, R.: 'Robust text-independent speaker identification using Gaussian mixture speaker models', *IEEE Trans. Speech Audio Process.*, 1995, **3**, (1), pp. 72–83
4 Esteller, R., Vachtsevanos, G., and Henry, T.: 'Fractal dimensions characterizes seizure onset in epileptic patients'. IEEE Proc., Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP99, Phoenix, AZ, USA, 1999, Vol. 4, pp. 2343–2346
5 Morimoto, T., *et al.*: 'Pattern recognition of fruit shape based on the concept of chaos and neural networks', *Comput. Electron. Agric.*, 2000, **26**, pp. 171–186
6 Fernández, S., Feijóo, S., and Balsa, R.: 'Fractal characterization of spanish fricatives'. Proc. ICPhS, 1999, pp. 2145–2148
7 Petry, A., and Barone, D.: 'Fractal dimension applied to speaker identification'. Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP, San Francisco, CA, USA, 2001
8 Veith, D., and Abry, P.: 'A wavelet-based joint estimator of the parameters of long-range dependence', *IEEE Trans. Inf. Theory*, 1998, **45**, (3), pp. 878–897
9 Daubechies, I.: 'Ten lectures on wavelets' (SIAM, Philadelphia, 1992)
10 Martin, A., *et al.*: 'The det curve in assessment of detection task performance'. Proc. EuroSpeech 97, Rhodos, Greece, 1997, pp. 1895–1898
11 Reynolds, D., Rose, R., and Hofstetter, E.: 'Integrated models of signal and background with application to speaker identification in noise', *IEEE Trans. Speech Audio Process.*, 1994, **2**, pp. 245–267