FRACTAL DIMENSION APPLIED TO SPEAKER IDENTIFICATION

Petry, A. and Barone, D. A. C.
{adpetry,barone}@inf.ufrgs.br

Instituto de Informática, Universidade Federal do Rio Grande do Sul - Porto Alegre, Brazil

ABSTRACT

This paper reports the results obtained in a speaker identification system based in Bhattacharrya distance, which combines LP-derived cepstral coefficients, with a nonlinear dynamic feature namely fractal dimension. The nonlinear dynamic analysis starts with the phase espace reconstruction, and the fractal dimension of the correspondent attractor trajectory is estimated. This analysis is performed in every speech window, providing a measure of a time-dependent fractal dimension. The corpus used in the tests is composed by 37 different speakers, and the best results are obtained when the fractal dimension is included, suggesting that the information added with this feature was not present so far.

1. INTRODUCTION

A speaker identity is strongly dependent of the physiological and behavioral characteristics of the speech production system. The first step of a basic speaker recognition system is to extract from the speech samples a "good" parametric representation. These parameters must be, as much as possible, representative of a speaker, presenting low variability for that speaker's speech samples, and great difference when used with others speakers' speech samples.

Previous papers [1][2][3][4] have worked with speech characterization and analysis using nonlinear dynamical features. Sabanal *et al*. [2] used the time-dependent fractal dimensions (TDFDs), extracted through critical exponent method (CEM), and the time-dependent multifractal dimensions (TDMFDs) to accomplish a speech recognizer. The target was to recognize Japanese digits using a neural network. Kumar *et al*. [1] estimated Lyapunov exponents, dimension and metric entropy in phonemes signals, divided into eight different types. Banbrook *et al*. [3] extracted correlation dimension, Lyapunov exponents, and short-term predictability from a corpus of sustained vowels sounds. The works mentioned used some nonlinear dynamical features to characterize a speech sound, showing the speech low dimensionality and the average exponential divergence of nearby trajectories in the reconstructed phase space.

In this work a speaker identification is performed using a combination of LP-derived cepstral coefficients with a nonlinear dynamic invariant: the fractal dimension. While the use of LP-derived cepstral coefficients can perform a speaker recognition quite successfully, it may not be accurate enough for some applications. The characterization of a speaker using a nonlinear dynamic description can help on identifying people from their voices. The assumptions used to extract the standard feature parameters do not describe the nonlinear dynamic evolution of the system. It will be shown that add nonlinear dynamic qualitative information to the standard feature parameters, such as fractal dimension, is equivalent to add speaker-dependent features, not present in the standard feature parameters so far. This combination will lead a speaker recognition system to more accurate results.

2. PHASE ESPACE RECONSTRUCTION

In experimental applications, it is often available unidimensional measurements of a dynamical system that evolves in a multidimensional phase space. This scalar time series contains the information available from that system. In many cases, no further information is available, and an important challenge that has to be solved is the calculation of the system's real multidimensional phase space trajectory. After that, measurements that provide important knowledge about the system behavior can be done.

To evaluate the properties of an attractor associated to a time series it is first necessary to reconstruct its evolution in a proper phase space. The most used way of reconstructing the full dynamics of a system from scalar time series measurements was proposed by Takens [5]. This method presents easy practical implementation. Given a *N*-point time series $x(t_i)$ for $i=1,2,...N$ as follows

$$x(t) = \{x(t_1), x(t_2),..., x(t_N)\},$$

the *m*-dimensional vectors are reconstructed, according to Takens delay method [5], as

$$\vec{X}_i = \{x(t_i), x(t_i + p), x(t_i + 2p),..., x(t_i + (m-1)p)\},$$

where *p* is called time delay and *m* is the embedding dimension. The $\vec{X}_i$ vectors represent the trajectory of the time series $x(t_i)$ in a *m*-dimensional phase space.

The choice of the proper time delay (*p*) and embedding dimension (*m*) values must be made carefully. A too small value to time delay produces vectors $\vec{X}_i$ and $\vec{X}_{i+1}$ very similar, and consequently an autocorrelated attractor trajectory, probably stretched along the diagonal. When *p* value is excessive the reconstructed trajectory becomes too disperse. If the attractor is unfolded into a phase space whose embedding dimension is lower than the minimum necessary, there will be vectors that remain close to one another not because of the system dynamics. On the other hand, if the chosen embedding dimension is too high, the number of vectors $\vec{X}_i$ is reduced, and it is a problem for time series composed by limited *N* numbers of points.

A criterion for an intermediate choice of time delay values is based on the analysis of autocorrelation function [6]. The autocorrelation function provides a measure of the similarity between the samples of a signal, and typically the value of *p* is set as the delay where the autocorrelation function first drops to half of the initial value. Other methods for choosing time delay can be found in [6].

An interesting method to estimate an acceptable minimum embedding dimension is called method of false neighbors [7]. Basically, for each vector of the reconstructed attractor trajectory, unfolded into a *d* embedding dimension phase space, a search for its nearest neighbor vector is made. When the embedding dimension is increased to *d+1*, it is possible to discover the percentage of neighbors that were actually "false" neighbors, and did not remain close because the *d* embedding dimension was too small. When the false neighbors percentage drops to an acceptable value, it is possible to state that the attractor was completely unfolded.

3. FRACTAL DIMENSION ESTIMATION

Nakagawa [8] estimated the fractal dimensions of self-affine data with power spectra in according with a power law based on the moment exponent. The theoretical outlines of this method, called critical exponent method (CEM), is reviewed below and a particular adaptation for speech signals is suggested.

For time series of self-affine data, the fractal dimension D_0 can be estimated as

$$D_0=2-H\ ,$$

where H is the Hurst exponent.

The CEM is based on analyzing the momentum I_α associated to the signal power spectrum, defined as

$$I_\alpha = \int_1^U duP(u)u^\alpha\ ,\ \text{for}\ -\infty<\alpha<+\infty\ ,$$

where U is the upper limit to the normalized frequency u, and $P(u)$ is the power spectral density, and may be assumed to follow the power law

$$P(u)\approx u^{-\beta}\ .$$

Specifically to speech signals, consider k_c the lower cut frequency below which $P(u)$ does not follow the power law. By making $u=k/kc$, where k is the real frequency, these low frequencies are correctly not considered [8]. However, the estimation of proper values for k_c is made heuristically, by visualizing the speech power spectrum and "guessing" the correct value. In this work, we suggest a way to determine automatically a good approximation to k_c based on a smoothness representation of the power spectral density.

Figure 1 shows a typical frequency response from a 30ms voiced speech window, obtained through fast Fourier transform (FFT) algorithm. It is important to note that, after a determined frequency, the power spectrum decreases, following approximately the power law described previously. If it is available a smoothness representation of the power spectrum, it would be possible to search for the maximum and consider it the lower cut frequency k_c, above which an approximately exponential decrease is presented. The use of a smoothness representation of frequency response, instead of only choosing the frequency whose magnitude is maximum in the window, considers all its evolution and avoid choosing a frequency based only in one (possibly incorrect) value. Furthermore, the smoothness representation is capable of providing the exact frequency where the magnitude stops increasing and starts decreasing, avoiding small peaks.

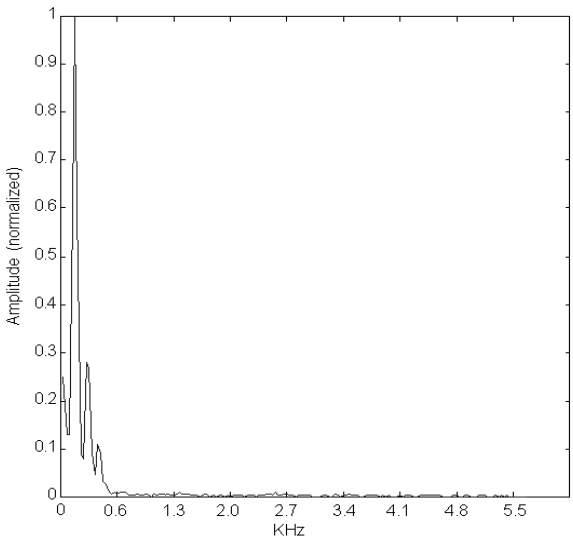

Fig. 1. Typical frequency response from a 30ms voiced speech window.

A smooth envelope representation of the power spectral density is available using the linear prediction spectrum [9][10], obtained from the estimation of the linear prediction coefficients (LPC). A simple peak picking algorithm can easily find a good approximation for k_c.

After determining the lower cut off frequency, for practical effects it is possible to differentiate the logarithm of moment I_α to the 3rd order using the following equation

$$\frac{\partial^3\ln(I_\alpha)}{\partial\alpha^3}=\frac{I_\alpha^2 I_\alpha^{'''}-3I_\alpha I_\alpha^{'}I_\alpha^{''}+2\left(I_\alpha^{'}\right)^3}{I_\alpha^3}\ ,$$

where the nth derivative of I_α, I_α^n, can be evaluated from the equation

$$I_\alpha^n=\frac{\partial^n}{\partial\alpha^n}\int_1^U duu^\alpha P(u)=\int_1^U du(\ln(u))^n u^\alpha P(u)\ ,$$

and solve the following equation to find the critical value α_c,

$$\frac{\partial^3\ln(I_\alpha)}{\partial\alpha^3}=0\,.$$

From the above relation, exponent β (from power law equation) is given as

$$\beta=\alpha_c+1=2H+1$$

and the fractal dimension D_0 can be calculated from

$$D_0=2-H=2-\frac{\alpha_c}{2}\,.$$

The method described previously can be applied in every speech window, which provides the time-dependent fractal dimensions (TDFDs). The values of fractal dimension obtained with automatic search for lower cut frequency k_c are very close to the results reported in [2] for speech signals. Figure 2 (b) shows the values of fractal dimensions obtained from the speech

signal in figure 2 (a), evaluating a 30ms hamming window of speech, applied every 10ms.

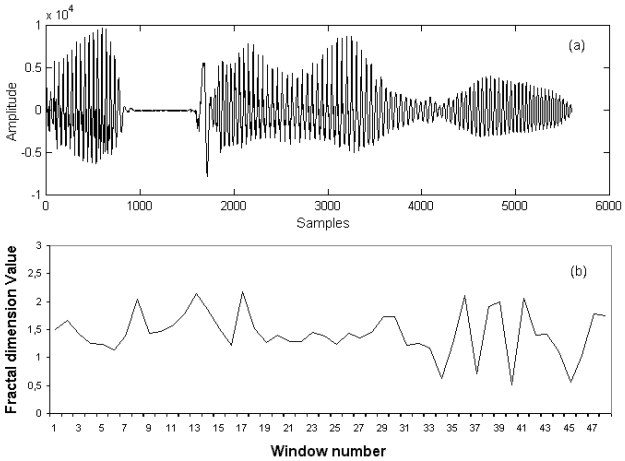

Fig. 2. Waveform of a speech signal (a), the respective time-dependent fractal dimensions (b).

4. THE SPEAKER RECOGNITION SYSTEM

A system based on the Bhattacharyya distance was used to evaluate the recognition performance that can be obtained when LP-derived cepstral coefficients is combined with fractal dimension. This system is similar to the one described in [11]. Basically, some speech samples of every registered speaker in the system are used to compose that speaker identity. It is done by extracting the desired feature parameters from every window of all speech samples from that speaker, and calculating its mean and covariance matrix. When an unknown speech sample is presented to the system, its feature parameters are extracted the same way, the mean and covariance matrix are calculated and a similarity measure is obtained for every registered speaker, using the Bhattacharyya distance for multivariate Gaussian distributions. The unknown speech sample is then assigned to the registered speaker whose similarity measure is maximized.

4.1. LP-derived cepstral coefficients

Linear prediction (LP) analysis is an important method of characterizing the spectral properties of speech in the time domain. In this analysis method, each sample of the speech signal is predicted as a linear weighted sum of the past *p* samples. The weights which minimize the mean-squared prediction error are called the predictor coefficients. The value of *p* is approximately determined by the number of poles of the vocal tract and the glottal wave transfer function, mathematically modeled. An important method to estimate the linear prediction coefficients (LPC) is called Durbin method, well detailed in [9][10].

By definition, the cepstrum (or the cepstral coefficients) is the inverse Fourier transform of the logarithm of the speech signal spectrum. The cepstral coefficients obtained from the predictor coefficients are called LP-derived cepstral coefficients. The relationship between the cepstrum and the predictor coefficients are [9][10]:

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \quad 1 \le m \le p$$

$$c_m = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \qquad m > p$$

where c_m is the *m*th cepstral coefficient, a_m is the *m*th linear prediction coefficient and *p* is the predictor order.

4.2. Bhattacharyya distance

In statistics, the proximity degree between two different probability densities is related with the notion of distance measure. An estimation to the upper bound on the Bays error can be obtained using the Bhattacharyya distance. Considering two probability densities $p_1(x)$ and $p_2(x)$, obtained from two different classes of feature parameters, the Bhattacharyya distance [12] is defined by

$$B = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)\,p_2(x)}\,dx$$

Special cases of this general distance measure can be calculated explicitly to a large types of probability densities. An important case refers to the multivariate Gaussian distributions. Considering $p_i(x)$ Gaussian probability densities, it is possible to show [13] that the previous equation can be written as:

$$B = \frac{1}{8}(m_1 - m_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1}(m_1 - m_2) + \frac{1}{2}\ln\left(\frac{\det(\Sigma_1 + \Sigma_2)/2}{\sqrt{\det(\Sigma_1)}\sqrt{\det(\Sigma_2)}}\right)$$

where m_i is the mean value and Σ_i is the covariance matrix, obtained from the feature parameters of class *i*.

The Bhattacharyya distance can be applied to a wide variety of known probability distributions, according to the best fit. The assumption of Gaussian density for the parameters is not arbitrary, since it is sufficient that the density be essentially unimodal and approximately Gaussian in the center of its range. These properties are often respected in physical systems. Inspecting histograms obtained from the feature parameters, it is possible to verify that their value distributions can be modeled as Gaussian probability densities.

4.3. The Data Set

The corpus used to evaluate the speaker recognition performance is composed with speech samples from 37 different speakers, sampled at 11025Hz, with resolution of 16 bits per sample. Every speaker provided three repetitions of the vocabulary, composed by the words “first”, “second”, ... , and “tenth”, spoken in Portuguese language. Every speech sample was about 618ms long, in average. The first two repetitions of the vocabulary were used to generate the speakers identity. The third repetition of the vocabulary of every speaker was used to test the system accuracy, in a total of 370 different identifications for a single test.

5. EXPERIMENTAL RESULTS

Different tests were accomplished, and the focus was to verify the efficiency of fractal dimension in speaker recognition task. From all speech samples, a hamming window with length of

30ms was applied every 10ms, and from every window LP-derived cepstral coefficients and fractal dimension were extracted according with the methods previously described.

The nonlinear feature parameters can help on improving the accuracy obtained with Fourier and cepstral analysis, by providing other kind of information, not considered so far. The combination of Fourier and cepstral, with nonlinear dynamic analysis can more accurately characterize a speaker, leading the correspondent speaker recognition system to higher performance. It can be seen in figure 3, where the number of LP-derived cepstral coefficients vary. When nonlinear dynamic information is combined with cepstral information there is a improvement in the system's performance, indicating that it contains speaker-dependent information, which can distinguish different speakers. The combination of cepstral analysis with nonlinear features leads the speaker identification system to even better results, achieving 97.29% of accuracy, which is a good result considering the amount of speech used on trainning (about 1.2s per speaker, in average) and recognition (about 618ms per identification, in average).

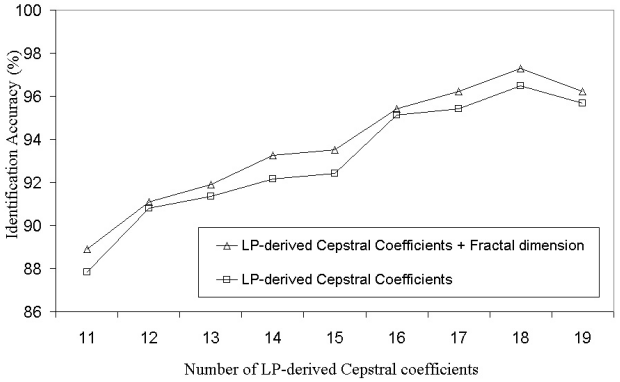

Fig. 3. System identification accuracy varying the number of LP-derived Cepstral coefficients.

Figure 3 shows clearly that there is a considerable performance gain when combining nonlinear dynamic features. However, the processing time necessary to extract the nonlinear features may be much greater than to extract cepstrum. For comparison, a personal computer with an Pentium processor running at 350 MHz takes about 69.2ms to extract fractal dimension, and only 6.9ms to extract 17 LP-derived cepstral coefficients from a window of 30ms of speech. The processing time is heavily increased when the nonlinear feature is added. It takes about 90.9% of the total processing time to complete the nonlinear dynamical analysis and only 9,07% of the total processing time for LP-derived cepstral coefficients extraction. The previous estimation of time was based in an average, using 289 different windows from a speech file.

6. CONCLUSIONS

This work suggests new ideas to construct speaker recognition systems more robust and reliable. Extract new information that specifically distinguish different speakers is very important to continue the development of this area. In the other hand, the introduction of new techniques and new features to characterize a speaker will bring an intrinsic computational processing overhead. Particularly with nonlinear features, such processing may not allow the construction of real time systems with the hardware available today.

Many applications where the speaker recognition technology can potentially be introduced are still searching for more accurate systems. The nonlinear dynamic analysis can analyze the speech production differently, as the result of a nonlinear dynamic process, bringing up new information to characterize it in a more complete way.

Acknowledgments – Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) and Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq).

7. REFERENCES

[1] Kumar, A., Mullick, S. K. Nonlinear dynamical analysis of speech, *J. Acoust. Soc. Am.* 100 (1), July 1996.

[2] Sabanal, S., Nakagawa, M. The Fractal Properties of Vocal Sounds and Their Application in the Speech Recognition Model, *Chaos, Solitons & Fractals*, Vol. 7, No. 11, pp. 1825-1843, 1996.

[3] Banbrook, M., McLaughlin, S. Mann, I. Speech Characterization and Synthesis by Nonlinear Methods, *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 1, January 1999.

[4] Chan, A. M. and Leung, H. Equalization of Speech and Audio Signals Using a Nonlinear Dynamical Approach, *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 3, May 1999.

[5] Takens, F. Detecting strange attractors in turbulence *in Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, edited by D. A. Rand and L. S. Young, Springer-Verlag, Berlin, 1981, vol. 898, pp. 366-381.

[6] Rosenstein, M. T., Collins, J. J. and De Luca, C. J. Reconstruction expansion as a geometry-based framework for choosing proper delay times, *Physica D* 73 (1994) 82-98.

[7] Kennel, M. B., Brown, R. and Abarbanel, H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction, *Physical Review A*, vol. 45, no. 6, pp. 3403-3411, March 1992.

[8] Nakagawa, M. A Critical Exponent Method to Evaluate Fractal Dimensions of Self-Affine Data, *J. of the Physical Society of Japan*, vol. 62, no. 12, December 1993.

[9] Deller Jr., J. R., Proakis, J. G. and Hansen, J. H. L. Discrete-time processing of speech signals, *Prentice Hall*, 1987.

[10] Rabiner, L. R. and Schafer, R. W. Digital processing of speech signals, *Prentice Hall*, 1978.

[11] Campbell Jr., J. P. Speaker Recognition: A Tutorial, Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-1462, September 1997.

[12] Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions, *Bull Calcutta Math. Soc*., vol. 35, pp. 99-109, 1943.

[13] Kailath, T. The Divergence and Bhattacharyya Distance Measures in Signal Selection, *IEEE Trans. On Communication Technology*, vol. Com-15, pp. 52-60, February 1967.