TEXT-DEPENDENT SPEAKER VERIFICATION USING LYAPUNOV EXPONENTS

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

The characterization of a speech signal using nonlinear dynamical features has been focus of intense research lately. In this work, the results obtained with time-dependent largest Lyapunov exponents (TDLEs) in a text-dependent speaker verification task are reported. The baseline system used 10 cepstral coefficients and 10 delta cepstral coefficients, and it is shown how the addition of TDLEs can improve the system’s accuracy. Cepstral mean subtraction (CMS) was applied to all features in the tests, as well as silence removal. The telephone speech corpus used, obtained from a subset of CSLU Speaker Recognition corpus, was composed by 91 different speakers, speaking the same sentence.

1. INTRODUCTION

Many physical phenomena present a complex behavior with fluctuations over time. Biological signals, as electroencephalograms (EEGs), electrocardiograms (ECGs), vocal sounds, and measures of arterial blood pressure, represent a great challenge related to its analysis and modeling. A detailed model of vocal tract should consider the time variation of vocal tract shape, the vocal tract resonances, losses due to heat conduction and viscous friction at the vocal tract walls, nasal cavity coupling, softness of the vocal tract walls, the effect of subglottal coupling with vocal tract resonant structure and radiation of sound at the lips [1]. A time-varying linear filter can model the effects of some of these factors, but others are very difficult to model. Some techniques have been proposed in the literature to analyze the non-linearities of dynamical systems, including those systems where it is not currently possible to merge with a mathematical model. They constitute the called Chaos theory or nonlinear dynamical systems theory.

The nonlinear dynamical systems theory can use time series to characterize the dynamical properties of the corresponding system, and extract information from these data. Thus, the speech production can be analyzed by the techniques behind this theory, and the information extracted can be applied to improve the accuracy of many speech processing systems, such as speaker recognition systems. Previous papers [2-9] have worked with speech characterization and analysis using nonlinear dynamical features.

We will explore in this paper the extraction of an important nonlinear dynamical feature, namely largest Lyapunov exponent, from every window of speech signals. This kind of analysis will provide a set of values, in this work referred to as time-dependent largest Lyapunov exponents (TDLEs). Furthermore, we intend to demonstrate that this feature can be successfully used in speech recorded at low sample rates (such as telephone speech), and it contains speaker-dependent information, which can improve the accuracy of a speaker recognition system.

2. STATE SPACE RECONSTRUCTION

In experimental applications, it is often available a set of one-dimensional measurements of a dynamical system that evolves in a multidimensional state 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 state 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 state space. The most common way of reconstructing the full dynamics of a system from scalar time series measurements was proposed by Takens [10]. This method presents an easy and practical implementation. Given a N-point time series \( x(t) \) for \( i=1,2,...,N \) as follows

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

(1)

the m-dimensional vectors are reconstructed, according to Takens delay method [10], as

\[
\hat{X}_i = \{ x(t_i), x(t_i + p), x(t_i + 2p), ..., x(t_i + (m-1)p) \}.
\]

(2)

where \( p \) is called time delay and \( m \) is the embedding dimension. The \( \hat{X}_i \) vectors represent the trajectory of the time series \( x(t) \) in a \( m \)-dimensional state space.

The choice of the proper time delay (\( p \)) and embedding dimension (\( m \)) values must be made carefully. An excessive small value assigned to time delay produces very similar vectors \( \hat{X}_i \) and \( \hat{X}_{i+1} \), and consequently an autocorrelated attractor trajectory, probably stretched along the diagonal. When \( p \) is too large, the reconstructed trajectory becomes too disperse. If the attractor is unfolded into a state 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 \( \hat{X}_i \) 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 the autocorrelation function. 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 [11].

An interesting method to estimate an acceptable minimum embedding dimension is called method of false neighbors [12]. 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. LARGEST LYAPUNOV EXPONENT ESTIMATION

Rosenstein et al. [13] proposed a method to estimate the Lyapunov exponents from time series composed by few samples. Good results were obtained for the largest Lyapunov exponent ($\lambda_1$) estimation of known systems using less than 1000 samples. This characteristic is very important when dealing with speech, since a speech signal can be considered stationary only during a small window of approximate 30ms [1]. Furthermore, it allows the correct estimation of Lyapunov exponents from speech windows, using speech recorded at low sample rates, such as telephone speech.

Rosenstein’s method for largest Lyapunov estimation is outlined as follows: the first step is the reconstruction of the attractor’s trajectory in an appropriate state space. After, the nearest neighbor of every vector of the reconstructed trajectory is found. A constraint that nearest neighbors have temporal separations greater than the mean period of the time series must be satisfied. Doing this, it is possible to consider the pair of neighbors as belonging to different trajectories. When considering two trajectories whose initial conditions are very similar, the trajectories diverge, on average, at an exponential rate characterized by the largest Lyapunov exponent ($\lambda_1$):

$$d_j(i) = C_j e^{\lambda_1 i \Delta t}$$

where $d_j(i)$ is the distance between the $j$th pair of nearest neighbors after $i$ steps (equals to $i \cdot \Delta t$ seconds, where $\Delta t$ is the time series sampling period) and $C_j$ is the initial separation between the neighbors.

When the natural logarithm is applied to both sides, the previous equation becomes

$$\ln d_j(i) = \ln C_j + \lambda_1 (i \Delta t)$$

If the logarithm of the distance evolution between every pair of neighbors is monitored, it will appear as a set of approximately parallel lines, each with a slope proportional to $\lambda_1$. The mean line, calculated from these parallel lines, can be best modeled by applying a least-squares method, and the largest Lyapunov exponent is then estimated as the modeled line slope. Figure 1 shows an example of the logarithm of the mean distance evolution between every pair of neighbors. The $n$-dimensional vectors were calculated from the state space reconstruction of a 30ms hamming window, obtained from the telephone speech signal in figure 2. It is easy to verify its positive slope, which indicates a positive value for the correspondent largest Lyapunov exponent.

The speech signal showed in figure 2 was extracted from CSLU Speaker Recognition Corpus. The complete information about this corpus can be found in [14]. The speech file of figure 2 was named as “00809al1.wav” in the corpus. By repeating the same process described above in the speech signal in figure 2 for every window of length of 30ms, applied every 10 ms, the TDLEs can be obtained and are showed in figure 3.

4. BASELINE SYSTEM

A system based on the Bhattacharyya distance was used to evaluate the recognition performance that can be obtained when LP-derived cepstral coefficients are combined with TDLEs. The baseline system in this work used 10 LP-derived cepstral coefficients added to 10 delta cepstral coefficients. Detailed information about cepstral coefficients estimation can be found in [1]. Cepstral mean subtraction (CMS) [15-16] was applied to the features for reduction of channel distortions, and silence frames were discarded based on energy estimation.
Basically, some speech samples of every registered speaker in the system are used to compose the speaker’s voiceprint. It is done by extracting the desired feature parameters from every window of all speech samples from that speaker, and calculating its mean vector 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 distortion measure can be obtained for any registered speaker, using the Bhattacharyya distance for multivariate Gaussian distributions [17]. The functionality of this system is similar to the one described in [18].

Speaker recognition can be classified into speaker identification and speaker verification. In speaker identification task, the unknown speech sample is compared with all registered speaker’s voiceprint. In this case, the unknown speech sample is assigned to the registered speaker whose similarity measure is maximized. For speaker verification, the claimed registered speaker’s identity must be also provided. Then, the unknown speech sample is compared only with the claimed registered speaker’s voiceprint. If the distortion measure obtained is lower than a threshold, the identity is accepted; otherwise, it is rejected by the system. The system used in this work tested speaker verification in a text-dependent task, and its accuracy was compared when TDLEs were added to the baseline features.

4.1 Bhattacharyya distance

In statistics, the proximity degree between two different probability densities is related with the notion of distance measure. One way to calculate this distance between classes is 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 [19] is defined as

$$B = -\ln \int \sqrt{p_1(x)p_2(x)}dx \quad (5)$$

Special cases of this general distance measure can be calculated explicitly for large types of probability densities. An important case refers to the multivariate Gaussian distributions. Considering $p(x)$ Gaussian probability densities, it is possible to show [20] that the previous equation can be written as:

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

$$+ \frac{1}{2} \ln \left( \frac{\det(\Sigma_1 + \Sigma_2)}/2}{\sqrt{\det(\Sigma_1)\sqrt{\det(\Sigma_2)}}} \right) \quad (6)$$

where $m_i$ is the mean vector and $\Sigma_i$ 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 of correspondent probability densities. 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 the histograms obtained from feature parameters, it is possible to verify that their distributions may be modeled as Gaussian probability densities.

4.2 The speech corpus

The speech corpus used in this work is a subset of CSLU Speaker Recognition corpus. The speakers’ speech samples were recorded from digital telephone lines, in various sessions during a two-year period. The sample rate used was 8000 Hz, with resolution of 16 bits per sample. Different transducers were used in the recordings, providing an unmatched condition. In this corpus, every speaker was prompted to repeat the same words and sentences as well as speak freely, allowing the tests of both text-dependent and text-independent systems. Detailed information about CSLU Speaker Recognition corpus can be found in [14].

The subset of the speech corpus used in this work was composed by speech samples from 91 different speakers. All the speakers provided 11 or 12 repetitions of speech samples with the same text: “If it doesn’t matter who wins, why do we keep score?” The speech samples whose information did not correspond to the complete expected sentence were discarded. Every repetition of the sentence above had duration of 3.2 s in mean. For training, 5 repetitions were used to generate every speaker’s voiceprint and the others were used for testing the system. The testing speech samples from every speaker were used to measure false rejection (FR) rate when compared with that speaker’s voiceprint, and were used to measure false acceptance (FA) rate when compared with other speakers’ voiceprint. The tradeoff between these two measures is a function of the decision threshold. The plot of FA versus FR provides the ROC curves, a powerful way of analyzing the performance of a system.

5. EXPERIMENTAL RESULTS

The experimental tests accomplished used the baseline features, composed of 10 cepstral and 10 delta cepstral coefficients. The system accuracy was compared when TDLEs were added. The estimation of largest Lyapunov exponents using Rosenstein’s method [13] can provide reliable results with few data. However, considering the sample rate of 8000 Hz used in the speech corpora, a window of 45 ms was used, and 360 samples from the speech window showed to be enough for Lyapunov measures, what was tested and reported in figure 4.

![Figure 4: System performance using 45 ms windows](image-url)
The addition of TDLEs led the Bhattacharyya-based, text-dependent speaker verification system to better performance. Above 45 ms of window length, the assumption of stationarity becomes unreliable.

6. CONCLUSIONS
This work suggested new ideas to construct more robust and reliable speaker recognition systems. Extraction of new information that specifically distinguishes different speakers is very important to continue the development of this area. The standard feature parameters widely used to characterize a speaker can perform this task with relative success. However, some applications where the speaker recognition technology can potentially be introduced are still waiting for more accurate systems. The nonlinear dynamics analysis can see the speech production differently, as the result of a nonlinear dynamical process, bringing up new information to characterize it in a more complete way, and perhaps pointing out a way of improving the accuracy of speaker recognition systems.

Chaos theory has been focus of intense research lately, mainly because of the great development of hardware capabilities. Computational algorithms were developed and used in real-world time series, providing good results even when the number of samples available was not large. In this work, the estimation of TDLEs was possible and accurate using telephone speech when the window length of 45ms was used. The addition of new information led the speaker recognition system under study to a more complete way of analyzing the signal, bringing up new information to characterize it in a more complete way, and perhaps pointing out a way of improving the accuracy of speaker recognition systems.

Acknowledgments - The authors would like to acknowledge Financiadora de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support.

7. REFERENCES