Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Nonlinear dynamics characterization of emotional speech

Patricia Henríquez [a,*], Jesús B. Alonso [a], Miguel A. Ferrer [a,b], Carlos M. Travieso [a], Juan R. Orozco-Arroyave [b]

[a] *Instituto Universitario para el Desarrollo Tecnológico y la Innovación en Comunicaciones (IDeTIC), Universidad de Las Palmas de Gran Canaria, 35017, Las Palmas de Gran Canaria, Spain*

[b] *Departamento de Ingeniería Electrónica, Universidad de Antioquia, GEPAR and GITA Research Groups, Medellín, Colombia*

ARTICLE INFO

Article history:
Received 22 February 2012
Accepted 4 May 2012
Communicated by the Chennai Guest Editor
Available online 21 October 2013

Keywords:
Nonlinear dynamics
Complexity measures
Emotional speech

ABSTRACT

This paper proposes the application of complexity measures based on nonlinear dynamics for emotional speech characterization. Measures such as mutual information, dimension correlation, entropy correlation, Shannon entropy, Lempel–Ziv complexity and Hurst exponent are extracted from the samples of three databases of emotional speech. Then, statistics such as mean, standard deviation, skewness and kurtosis are applied on the extracted measures. Experiments were conducted on the Polish emotional speech database, on the Berlin emotional speech database and on the LCD emotional database for a three-class problem (neutral, fear and anger emotional states). A procedure for feature selection is proposed based on an affinity analysis of the features. This feature selection procedure is accomplished to select a reduced number of features over the Polish emotional database. Finally, the selected features are evaluated in the Berlin emotional speech database and in the LDC emotional database using a neural network classifier in order to assess the usefulness of the selected features. Global success rates of 72.28%, 75.4% and 80.75%, were obtained for the Polish emotional speech database, the Berlin emotional speech database and the LDC emotional speech database respectively.

1. Introduction

Speech is one of the main modes of communication between human beings and an effective way to express emotions. Automatic recognition of human emotions in speech aims at automatically detecting the speaker emotional state based on speech and has attracted the research community in the last few years due to its applications in industry. Emotion recognition systems open new horizons in artificial intelligence with the improvement of voice synthesizer. The addition of emotional speech in voice synthesizer produces more natural speech facilitating human-computer interaction or computer aided communications [1]. Another application is the automatic recognition of negative emotions (e.g. anger or rage) in call centers based on interactive-voice-response systems. It is useful to detect problems in the customer-system interaction to help the customer by offering a human operator, for example [2–4]. In surveillance applications, speech emotion recognition can take an important role in detecting security threats using word-spotting techniques with the combination of emotional recognition.

One of the key steps in automatic recognition of emotional speech is the extraction of features that effectively distinguish between the emotions to be classified. In emotion recognition literature, features can be divided in linguistic and acoustic features. Linguistic features focus on explicit linguistic message (dialog related features) and acoustic features focus on implicit message. Acoustic features include prosodic features (which are mostly related to pitch, energy and speaking rate) [5], spectral and cepstral features such as Mel Frequency Cepstral Coefficients [5,6] and voice quality features such as harmonics-to-noise ratio, jitter or shimmer [7]. Features are usually extracted in a short-term basis, obtaining different number of features for each sample in the emotional database. Statistics variations of these features (functionals), such as their average, skewness, minimum, maximum and standard deviation are also frequently used. The use of functionals is probably justified by the supra-segmental nature of the phenomena found in emotional speech [7].

The success rates achieved in different works on automatic recognition of emotional speech are very difficult to compare due to the lack of a freely available corpus of reference and the lack of a standard methodology. Another issue is the different ways of considering emotions: as categorical discrete emotions or as continuous emotions in a multidimensional space (i.e. activation,

[*] Corresponding author. Tel.: +34 928459485; fax: +34 928400040.
E-mail addresses: phenriquez@gi.ulpgc.es (P. Henríquez), jalonso@dsc.ulpgc.es (J.B. Alonso), mferrer@dsc.ulpgc.es (M.A. Ferrer), ctravieso@dsc.ulpgc.es (C.M. Travieso), rafael.orozco@udea.edu.co (J.R. Orozco-Arroyave).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.05.037

valence, etc.). In the next few lines, we show examples of recent results obtained in the literature: using the Berlin emotional speech database [8] and classifying between 7 discrete emotions with both modulation spectral features and prosodic features a 91.6% of success rate is achieved [9]. A success rate of 88.6% is a obtained using spectro-temporal features and prosodic features [5] using the same database [8] and classifying between 7 discrete emotions as well. In another work, a maximum of 79% of accuracy is obtained with a set of linguistic and acoustic features in the classification between anger and no-anger speech using three different databases of real emotions [4].

Chaos theory is an area of nonlinear dynamics systems theory and has been adopted as a nonlinear approach to speech signal processing in the last two decades. Complexity features studied in the speech processing literature include dimension correlation, Rényi entropies [10], Lyapunov exponents [11], Hurst exponent [12] and Lempel–Ziv complexity [13]. These features have proved to be useful in distinguishing between different voice quality and in voice pathology detection [10–13]. In this paper, we propose the study of a set of nonlinear measures or complexity measures in emotional speech to assess its discrimination ability between neutral state, fear emotional state and anger emotional state produced by actors in three different emotional speech database: the Polish emotional speech database [24], the Berlin emotional speech database [9] and the Emotional Prosody Speech and Transcripts of the Linguistic Data Consortium (LCD) [25]. The study of the discrimination ability between neutral and negative emotional states such as fear or anger is very useful in surveillance applications to detect security threats. Recently, the authors proposed the same set of complexity measures using the Berlin emotional database [14]. In this paper, we extent the experimentation to the LCD database and to the Polish database to assess the discrimination ability of the proposed features and to compare the results. The complexity measures extracted are: the value of the first minimum of the mutual information function (MI), the Shannon entropy (SE), the Token’s estimator of the correlation dimension (CD), the correlation entropy (CE), the Lempel–Ziv complexity (LZC) and the Hurst exponent (H). Mean (μ), standard deviation (σ), skewness (sk) and kurtosis (k) are applied to the extracted measures, obtaining 24 features.

This paper also proposes a procedure to select a group of global optimal features. The selection is based on an affinity analysis of the previously selected features using the standard feature selection algorithm called Sequential Floating Forward Selection procedure [22]. The feature selection procedure is accomplished over the Polish emotional speech database to select a reduced number of features. Then, the selected features are extracted from the Berlin emotional database and from the LCD database. Finally, a neural network classifier is used to quantify the discrimination ability of the selected features in the three emotional databases.

The remainder of this paper is organized as follows: complexity features are described in Sections 2 and 3 is devoted to the speech emotional databases used in this paper, the experimental procedure is shown in Section 4. Section 5 shows the results and a discussion of the results. Finally, Section 6 is devoted to the conclusions.

2. Material and methods

2.1. Nonlinear dynamics systems: Taken's embedding

Deterministic dynamical systems describe the time evolution of a system in some phase space $\Gamma \in \Re^m$ (m-dimensional vectorial space), where a state is specified by a vector. The evolution in time can be expressed by ordinary differential equations or by maps in discrete time. The dynamical system underlying the speech production process is very complex and its equations are unknown. Nevertheless, Takens' embedding theorem [15] establishes that it is possible to reconstruct a phase space diffeomorphically equivalent to the original one from the time series of a system. The delays method is used to reconstruct the state-space vector formed by time-delayed samples of the observation (the speech signal)

$$\overline{s}_n = [s[n], s[n-\tau], \ldots, s[n-(m-1)\tau]] \tag{1}$$

where $s[n]$ is the speech signal, m is the minimum embedding dimension of the phase space reconstructed and τ is the time delay. The speech signal is embedded in the reconstructed phase space and its long-term evolution in the reconstructed phase space is called attractor. τ can be determined by calculating the first minimum of mutual information [16] and the false neighbors method [17] can be used to estimate the minimum value of the embedding m.

2.2. Mutual information function

The mutual information function measures the mutual dependency between two variables. When these two variables are a discrete signal $s[n]$ and its delayed version $s[n+\tau]$, the mutual information function measures the quantity of information we already possess about the value of $s[n+\tau]$ if we know $s[n]$. The mutual information estimator reads as [16,18]

$$I(\tau) = \sum_{i,j} p_{i,j}(\tau) \ln \left[\frac{p_{ij}(\tau)}{p_i p_j(\tau)} \right] \tag{2}$$

where p_i is the probability of finding a value of $s[n]$ inside the ith bin of the data histogram and $p_{ij}(\tau)$ the joint probability that $s[n]$ is in bin i and $s[n+\tau]$ in bin j.

The first minimum of the mutual information function marks the delay where mutual information adds maximal information to the knowledge we have from $s[n]$. The value of the mutual information function in the first minimum (MI) quantifies the degree of irregular behavior of a time series in the time of maximum difference of a signal with its delayed version.

2.3. Correlation dimension

The correlation dimension gives an idea of the complexity of the dynamics. More complex systems have a higher correlation dimension. In random processes, correlation dimension is not bounded, while in deterministic systems there tends to be a finite value and it can be a non integer number (fractal dimension). CD is given as [18]

$$D_2 = \frac{d \ln C(\varepsilon, N)}{d \ln \varepsilon} \cong \lim_{\Delta \ln \varepsilon \to 0} \frac{\Delta \ln C(\varepsilon, N)}{\Delta \ln \varepsilon} \tag{3}$$

with $C(\varepsilon,N)$ being the correlation sum of a set of points $\overline{s}_n$ ($n=1,\ldots,N$) of the speech signal attractor in the reconstructed embedding.

$$C(\varepsilon, N) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \theta(\varepsilon - ||\overline{s}_i - \overline{s}_j||) \tag{4}$$

where $\theta(s)=0$ if $s \le 0$ and $\theta(s)=1$ if $s > 0$ which counts the number of points inside the sphere with radius ε around $\overline{s}_i$. $C(\varepsilon,N)$ is the average fraction of points within a distance of ε from any other point. D_2 is estimated by calculating the local slope of the curve ln $(C(\varepsilon))$ against $\ln(\varepsilon)$ when the curve has a plateau for different values of the embedding dimension.

In this paper, the Takens–Theiler estimator [26] of D_2 is computed (CD). Takens-Theiler estimator is a maximum-likelihood estimator of