Computers and Electrical Engineering 58 (2017) 215–226

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals

Shabnam Gholamdokht Firooz*, Farshad Almasganj, Yasser Shekofteh

Biomedical Engineering Department, Amirkabir University of Technology, Hafez Ave., P.O. Box 15875-4413, Tehran, Iran

ARTICLE INFO

Article history:
Received 9 September 2015
Revised 6 July 2016
Accepted 6 July 2016
Available online 12 July 2016

Keywords:
Automatic speech recognition
Mel-frequency cepstral coefficients
Reconstructed phase space
Recurrence plot
Two-dimensional wavelet transform

ABSTRACT

The spectral-based features, typically used in Automatic Speech Recognition (ASR) systems, reject the phase information of speech signals. Thus, employing extra features, in which the phase of the signal is not rejected, may fill this gap. Embedding the speech signal in the Reconstructed Phase Space (RPS) and then extracting some useful features from it, is a recently considered approach in this field. In this paper, we will follow this approach by evaluating some useful features from the Recurrence Plot (RP) of the embedded speech signals in the RPS; the proposed features are evaluated via applying a two-dimensional wavelet transform to the resulted RP diagrams. The proposed features are examined in an ASR task alone and in combination with the traditional Mel-Frequency Cepstral Coefficients (MFCC). For the second case, using English TIMIT corpus, 3.94% absolute classification accuracy improvement in the phoneme recognition accuracy rate, against using only the MFCC features is gained.

1. Introduction

In recent decades, a variety of linear models for speech coding, synthesis, and recognition with acceptable performances have been introduced. In this way, many types of research achieved improvement in the field of speech recognition by employing novel methods [1,2] or the detection of mispronunciation using Hidden Markov model [3]; however, there are nonlinear aerodynamic phenomena in the human speech production system which generally could not be included in linear models [4]. Therefore, nonlinear methods could potentially provide effective computational models to extract acoustic features which are useful for the nonlinear phenomena detection [4]. Furthermore, some recent studies have shown that utilizing nonlinear characteristics may improve the performances of the ASR systems [5].

Usual ASR systems exploit frequency domain features like Mel frequency cepstral coefficients [6]. The traditional frequency domain methods typically extract only the first and second order properties from the spectral patterns of speech signals [7]. However, there are many signals produced via nonlinear differential equations that have wide spectral characteristics [8]. In such cases, the frequency domain techniques are deficient, because it is impossible to dissociate the information of such a signal only in the frequency domain [5].

In addition, the speech signal shows some chaotic behaviors due to the existence of nonlinear phenomena such as the turbulence [9]. Studies about dynamical systems and chaos theory resulted in a kind of signal representation, a multi-dimensional trajectory embedded in the reconstructed phase space [4–6]. Following this approach, some useful features

* Corresponding author. Fax: +982166495655.
E-mail address: sh.firooz@aut.ac.ir (S.G. Firooz).

http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

were introduced, such as Lyapunov exponents (LE) and the Fractal Dimensions (FD) which investigate chaotic and nonlinear dynamical properties of the signals. Lyapunov exponents of the signals mapped to the phase space are evaluated to characterize a multi-dimensional chaotic time series [10]. In [11], the Fractal dimension of the speech signal and the MFCC features are simultaneously utilized in the ASR system. By adding these nonlinear features to the entire process, the experiment conducted on the Broadcast News ASR system in Spanish has shown 1.36% correct word rate (CWR) improvement against the baseline system which uses the MFCC features alone [11].

Recently, some non-classic feature extraction algorithms are directly developed over the multi-dimensional RPS transformation of the speech signal. The embedding process is performed using a set of input-output pairs reconstructed via the time delayed based approach. By selecting a sufficient dimension for the reconstructed space, the underlying theory guarantees that the dynamics of signals are fully accommodated in the attractors constructed by embedding them into the RPS. To extract beneficial parameters from the transformed signals, many studies in this field concentrated on the assessment of dynamical invariants [7,10] which do not rely on the initial conditions [4]. Some recent research have utilized statistical distributions such as the Gaussian Mixture Model (GMM) and applied it to signal trajectories appeared in the RPS [5-8]. In [8], the Poincaré section is employed as an effective tool to analyze the trajectories generated in the RPS, and a statistical modeling approach based on the GMM is applied to the Poincaré sections of the speech attractors to extract parameters which could help the ASR performance. Moreover, Ref. [9] applied a multi-dimensional linear method to model the speech trajectory reconstructed inside the RPS, using the Multivariate Autoregressive (MVAR) method, to improve the performance of the involved ASR system. In [12], a set of the Gaussian Mixture Models (GMMs) were trained over the phoneme attractors in the RPS, via which a proper feature vector could be evaluated; the posterior probabilities of different phonemes are then estimated by an MLP-based classifier. By applying this approach, 1.89% absolute accuracy rate improvement (for FARSDAT corpus) was gained, compared to the baseline system which used only the MFCC features.

As aforementioned, the reconstructed trajectories in the phase space could carry nonlinear dynamics of the involved systems; however, these high-dimensional trajectories could not be directly visualized. As the recurrence property is an essential feature of dynamical systems, it could be employed to investigate the system's behavior in the RPS domain. To follow this approach, the Recurrence Plot (RP) as a useful tool may be exploited to analyze the speech trajectories in the RPS [13]. In fact, the RP computes a binary square matrix to represent some main specifications of the phase space trajectories [13].

In this paper, a two-step approach is introduced: first, the speech signal is embedded in the phase space; next, it is parameterized using the recurrence property analysis of dynamical systems. In the final step, the RPS-based evaluated features resulted from the first step are combined with the common MFCC features to improve the extended continuous speech recognition (CSR) task.

1.1. Contribution

Investigating new efficient feature extraction methods to be applied to speech signals is an essential topic in the field of ASR tasks; the ultimate viewpoint of these efforts is basically to improve the performance of the designed and implemented systems. In this paper, we intend to follow and verify that the reconstructed phase space is a proper tool for capturing the signal nonlinear dynamics and compensating the phase information lack which occurs for common popular spectral features like the MFCC; however, the high computational demand of these features and the low accuracy condition that appears while evaluating some of them, makes this approach difficult or somewhere impractical for many real-time applications. To overcome these limitations existed for some previous proposed RPS-based techniques [12], this paper proposes an effective algorithm to extract proper auxiliary features from speech trajectories reconstructed in the phase space. The proposed method benefits from the feature extraction methods introduced in the speech and image fields, simultaneously. In this approach, the one-dimensional speech signal is converted into a two-dimensional image, as is dubbed the RP; the accuracy improvement of the ASR system is obtained by combining the features extracted from the RP-based images of the phone acoustic signals and the traditional MFCC spectral features promised for speech recognition purposes. Moreover, by conducting proper experiments, it is shown that the proposed features keep their performance in noisy conditions too; this is not the case for the features directly evaluated from the RPS domain, for they are typically sensitive to the initial conditions and the environmental noise [8].

The rest of this paper is organized as follows. Section 2 describes the embedding procedure that enables us to reconstruct the trajectory of a speech signal in the phase space. This section also gives a brief synopsis of the recurrence plot and the way it demonstrates some aspects of the dynamics of speech signals [13]. In the following, the proposed feature extraction methods applied to the RP images will be described in detail. Issues of dimensionality reduction based on the Linear Discriminant Analysis (LDA) and the Forward feature selection are also discussed in Section 2. The experimental methodology and data description are described in Section 3. The experimental results and discussion are presented in Section 4. The paper is concluded in Section 5.

2. Material and methods

As mentioned earlier, based on the RPS and RP theories, speech frames are converted into the RP-based images. Image processing methods are then used to extract proper features from the RP images. Since the wavelet transform has been used