Engineering Applications of Artificial Intelligence 44 (2015) 1–9

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

MLP-based isolated phoneme classification using likelihood features extracted from reconstructed phase space

Yasser Shekofteh [a,b,*], Farshad Almasganj [a], Ayoub Daliri [c,d]

[a] Biomedical Engineering Department, Amirkabir University of Technology, Tehran, Iran
[b] Research Center for Development of Advanced Technologies (RCDAT), Tehran, Iran
[c] Department of Speech and Hearing Sciences, University of Washington, Seattle, WA, USA
[d] Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA, USA

ARTICLE INFO

Article history:
Received 20 November 2014
Received in revised form
2 May 2015
Accepted 4 May 2015
Available online 22 May 2015

Keywords:
Isolated phoneme classification
Nonlinear speech processing
Phoneme attractor
Reconstructed phase space
Gaussian mixture models

ABSTRACT

Nonlinear properties of a complex signal can be represented in reconstructed phase space (RPS). Previously, researchers have developed RPS-based feature extraction approaches to capture nonlinear properties. Typically, these approaches are more computationally demanding – higher run-time – and less accurate than traditional techniques such as Mel-frequency cepstral coefficients (MFCCs) that fail to capture nonlinear properties of signals. To overcome these issues, we propose a new RPS-based feature extraction approach that is based on a previously reported approach. The proposed approach calculates the similarities between the embedded speech signals and a set of predefined speech attractor models in the RPS, and uses the similarities as a set of proper input features for a final phonetic classifier. A set of Gaussian mixture models (GMMs) is trained to represent the variety of all phoneme attractors in the RPS. Using the developed GMMs, for each embedded out-sample speech signal, a feature vector is calculated that consists of the Log-likelihoods. Then, an MLP-based classifier is used to estimate posterior probabilities for the phoneme classes. To test the performance of the proposed approach, we apply the approach to a Persian speech corpus (i.e., FARSDAT). Results show 1.89% absolute classification accuracy improvement in comparison to performance of a baseline system that exploits MFCC features. Combining different classifiers that use the proposed RPS-based features and MFCC features, the classifier gain the highest accuracy of 68.85% phoneme classification rate, with absolute accuracy improvements of 4.78% against a baseline system.

1. Introduction

Commonly, automatic speech recognition (ASR) systems use spectral-based features to model acoustic units of the speech signal. Optimal feature vectors have small dimensions (i.e., compact) and can discriminate between different acoustic units (Breslin, 2008). They need to cover all useful information needed to distinguish acoustic units (e.g., phonemes, syllables, or words) and to suppress irrelevant information of the speech signal that is not useful for speech recognition purpose. To select the optimal feature vectors, researchers use several feature extraction approaches such as linear prediction cepstral coefficients (LPCC), log filter-bank energy (LFBE), perceptual linear prediction (PLP), and Mel-frequency cepstral coefficients (MFCC) (Huang et al., 2001; Alam et al., 2013; Goh et al., 2014). These approaches are based on the assumption that the speech production system is a linear system; therefore, they do not take into account nonlinear properties of the speech signal—e.g., they remove the phase information of the speech signal (Indrebo et al., 2006; Loweimi et al., 2013). However, recent studies have improved the traditional feature extraction approaches by including nonlinear properties of the speech signal (Mowlaee et al., 2014; Drugman et al., 2014; Shekofteh and Almasganj, 2013a, 2013b; Kokkinos and Maragos, 2005; Paliwal and Alsteris, 2005).

Nonlinear properties of a complex signal can be represented in reconstructed phase space (RPS) (Kantz and Schreiber, 2004). This technique, which is based on the chaos theory, uses delay coordinate theorem to represent chaotic signals such as electrocardiogram (ECG), electroencephalogram (EEG), and the speech signal (Sharma and Pachori, 2015; Zhou et al., 2015; Koulaouzidis et al., 2015; Lee et al., 2014; Al-Fahoum and Qasaimeh, 2013; Thasleema et al., 2012; Nejadgholi et al., 2011; Johnson et al., 2005). Researchers have successfully implemented RPS-based feature extraction approaches in verity of speech applications (e.g., speech recognition, phoneme classification, pitch mark detection and speech enhancement) (Sun

[*] Corresponding author at: Biomedical Engineering Department, Amirkabir University of Technology, Hafez Avenue, PO Box 15875-4413, Tehran, Iran. Tel.: +98 21 64542372; fax: +98 21 66495655.
E-mail address: y_shekofteh@aut.ac.ir (Y. Shekofteh).

http://dx.doi.org/10.1016/j.engappai.2015.05.001
0952-1976/© 2015 Elsevier Ltd. All rights reserved.

et al., 2008, 2007; Povinelli et al., 2006; Hagmüller and Kubin, 2006). During mapping a signal from time domain to the RPS domain, proper adjustment of embedding parameters (such as time lag and embedding dimension) could result in RPS representations that topographically are equivalent with representations of the signal's generating system in the phase space (or state space) (Lang, 2002; Takens, 1981). Speech signal is a one-dimensional discrete time signal that is produced by a highly nonlinear and dynamic system (i.e., speech production system); by mapping this signal to the RPS domain, the nonlinear properties of the signal – and thus the nonlinearity of the underlying system – can be captured.

Initially, RPS-based techniques of speech processing were limited to dynamical features such as the Lyapunov exponents, correlation dimensions, box counting, and Higuchi-Katz fractal dimensions (Ezeiza et al., 2013; Pitsikalis and Maragos, 2009; Pitsikalis et al., 2003; Maragos and Potamianos, 1999). However, in more recent years, Povinelli et al. used parametric and nonparametric statistical procedures to model the attractors' structure of the representations of isolated phonemes in the RPS domain (Povinelli et al., 2006, 2004; Johnson et al., 2005). Furthermore, researchers have developed ASR techniques that are based on a combination of the popular MFCC features and RPS-based features of the speech signal (Jafari and Almasganj, 2012; Jafari et al., 2010). Although the RPS-based techniques capture the nonlinear properties of the signals, most of these techniques, to date, are more computationally demanding – higher run-time – and less accurate than popular techniques such as MFCC.

Overall, the capability of RPS-based techniques makes them unique for capturing the nonlinear properties of highly dynamic system such as speech production; however, their high computational demand and low accuracy makes them impractical for most applications. To overcome the limitations of previous RPS-based techniques (i.e., high computational demand and low accuracy), here, we proposed a novel RPS-based technique. This technique was developed on the basis of Povinelli's approach. We extracted the RPS-based features of the speech signal based on Log-likelihood score of optimally developed models of phoneme attractors in the RPS domain. Then, we combined these features using an MLP-based nonlinear posterior probability estimator. We used MLP-based classifiers given their higher accuracy in capturing highly dynamical features, as reported by several recent studies (Zamora-Martnez et al., 2014; Zarrouk et al., 2014; Tüske et al., 2012, 2011; Park et al., 2011; Valente et al., 2010). Then we compared the performance of MLP-based classifiers with several traditional classifiers. To achieve optimally developed models, we modelled the phoneme attractors using a full or diagonal covariance Gaussian Mixture Models (GMM) with several numbers of mixtures. We examined the performance of the technique in phoneme recognition based on accuracy and run-time indices. Our experimental results showed that (a) the performance of the proposed technique is better than previous RPS-based techniques and at the level of the performance MFCC, and (b) using MLP-based classifier in combination with RPS-based features results in higher accuracy in comparison to using other classifiers.

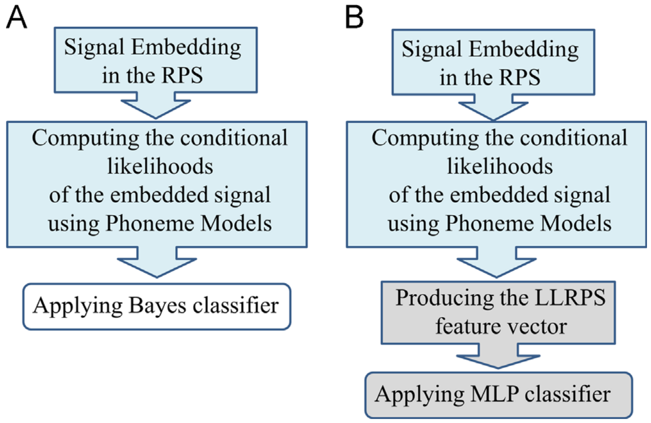
A
Signal Embedding in the RPS
Computing the conditional likelihoods of the embedded signal using Phoneme Models
Applying Bayes classifier
B
Signal Embedding in the RPS
Computing the conditional likelihoods of the embedded signal using Phoneme Models
Producing the LLRPS feature vector
Applying MLP classifier

Fig. 1. The block diagrams of two RPS-based phoneme classification approaches: (a) Povinelli's approach (Povinelli et al., 2006) and (b) the proposed approach in this study. In the proposed approach (b), the first two steps (light blue) were adopted from Povinelli's approach, and the last two steps (grey) were proposed to improve the approach. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The paper is organized as follows: Section 2 reviews previous RPS-based techniques of phoneme classification. Section 3 details the overall framework of the proposed classification method, including the feature extraction approach. In Section 4, we introduce the speech corpus and experimental setup. Section 5 introduces the experimental results. In Section 6, the combination of RPS-based and MFCC features are introduced and the obtained results are discussed. Finally, the discussion and conclusions are presented in Sections 7 and 8, respectively.

2. Phoneme classification

Given the capacity of RPS-based features in characterizing nonlinear aspects of signals, Povinelli et al. proposed the first RPS-based phoneme classification method that consisted of three steps (see Fig. 1, panel a): preprocessing and data transformation, modeling the densities of the phoneme signal trajectories in the RPS, and classifying the modeled densities by a Bayesian classifier. In the first step (i.e., preprocessing and transformation), two main analyses were conducted: (a) normalizing the original signal and (b) estimating the proper embedded dimension and time lag of the RPS. Normalization was done through subtraction of the mean of the signal and dividing by its standard deviation. The embedded dimension of the RPS was calculated based on the global false nearest-neighbor technique; the time lag (delay time) of the RPS was computed using the histogram mod of the first minimum of the automutual information function of all available signals (Kantz and Schreiber, 2004). In the second step (i.e., modeling the phoneme attractors), a GMM was developed for each embedded signal class in the RPS. GMMs were used to learn the structure of the geometric distributions of phoneme attractors. In the third step (i.e., classification), the developed GMMs in previous step in combination with a maximum likelihood Bayes classifier were used to classify the isolated phonemes. More specifically, this was accomplished by computing the conditional likelihoods of a test signal under different phoneme attractor models in the RPS and selecting the phoneme class with the highest likelihood.

To test this approach, Povinelli et al. computed point-by-point likelihood using the developed GMMs. Based on this approach, a vector with K elements (K was the number of phoneme classes) was created for a given test signal. Each element of the vector, which corresponded to a specific phoneme class, contained the averaged statistical scores for all samples of the test signal transformed to the RPS. Finally, a phoneme class corresponding to the element of the vector with the highest score was assigned as the phoneme class of the test signal.

However, Povinelli's approach is associated with two important drawbacks. First, this approach is very complex with high computational demand (i.e., high run-time). These issues may be resulted from the large number of Gaussian components in the GMMs that were used to build the phoneme attractors. Second, the classification scores for each class are only calculated and sorted to find the winner class. Given these drawbacks in Povinelli's technique, in this study, we propose a new approach that uses the first two steps of Povinelli's technique and two novel steps to overcome these issues (see Fig. 1, panel b).