Digital Signal Processing 49 (2016) 116–125

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Fuzzy decision fusion of complementary experts based on evolutionary cepstral coefficients for phoneme recognition

Seyed Mostafa Mirhassani [a,*], Hua Nong Ting [a], Abdorreza Alavi Gharahbagh [b]

[a] Biomedical Engineering Department, University of Malaya, Kuala Lumpur, Malaysia
[b] Department of Electrical and Computer Engineering, Islamic Azad University, Shahrood Branch, Shahrood, Iran

ARTICLE INFO

Article history:
Available online 19 November 2015

Keywords:
Evolutionary computation
Fuzzy decision fusion
Automatic speech recognition
Cepstral coefficients
Phoneme classification

ABSTRACT

Optimal representation of acoustic features is an ongoing challenge in automatic speech recognition research. As an initial step toward this purpose, optimization of filterbanks for the cepstral coefficient using evolutionary optimization methods is proposed in some approaches. However, the large number of optimization parameters required by a filterbank makes it difficult to guarantee that an individual optimized filterbank can provide the best representation for phoneme classification. Moreover, in many cases, a number of potential solutions are obtained. Each solution presents discrimination between specific groups of phonemes. In other words, each filterbank has its own particular advantage. Therefore, the aggregation of the discriminative information provided by filterbanks is demanding challenging task. In this study, the optimization of a number of complementary filterbanks is considered to provide a different representation of speech signals for phoneme classification using the hidden Markov model (HMM). Fuzzy information fusion is used to aggregate the decisions provided by HMMs. Fuzzy theory can effectively handle the uncertainties of classifiers trained with different representations of speech data. In this study, the output of the HMM classifiers of each expert is fused using a fuzzy decision fusion scheme. The decision fusion employed a global and local confidence measurement to formulate the reliability of each classifier based on both the global and local context when making overall decisions. Experiments were conducted based on clean and noisy phonetic samples. The proposed method outperformed conventional Mel frequency cepstral coefficients under both conditions in terms of overall phoneme classification accuracy. The fuzzy fusion scheme was shown to be capable of the aggregation of complementary information provided by each filterbank.

1. Introduction

The extraction of meaningful and discriminative information from speech signals, known as feature extraction, is the first step of most automatic speech recognition approaches. Aside from phonetic information, numerous sources of variability exist in any kind of recorded speech. These sources include background noise and speaker characteristics. Consequently, the extracted features should be robust to noise and reflect the phonetic identity of the speech. This goal can be achieved by enhancing the discrimination of phonetic information in speech signals. For this purpose, signal processing techniques, including filterbanks, cepstrum analysis, and linear prediction, are usually employed in most automatic speech recognition (ASR) approaches [1]. The Mel frequency cepstral coefficient (MFCC) is one of the most well-known feature extraction methods in ASR systems. This approach was originally introduced by Davis and Mermelstein (DM) [2]. Based on the psychoacoustic scale provided by the Mel filterbank, cepstral information is exploited for the subsequent steps of ASR systems. MFCC has been implemented in different platforms for speech recognition [3].

However, speech recognizer performance degrades in noisy environments. Thus, the extraction of a noise-robust feature is essential. The explanation on the Mel filterbank proposed by DM was used by subsequent researchers to propose different methods of providing MFCCs. The MFCCs introduced in Slaney's [4] Auditory Toolbox for Matlab doubled the number of Mel filters but reduced the spacing between the center frequencies. Additionally, wider frequency ranges were covered by each filter. In another modification provided by the Cambridge hidden Markov model toolkit (HTK) [5], the use of various design parameters, including the number of filters, frequency range, vocal tract length normalization, and use of signal magnitude or square magnitude, was made possible. The human factor cepstral coefficients [6,7] were

[*] Corresponding author. Fax: +60 379 674 579.
E-mail addresses: mostafamirhassani@gmail.com, s.m.mirhassani@siswa.um.edu.my (S.M. Mirhassani).

http://dx.doi.org/10.1016/j.dsp.2015.11.004
1051-2004/© 2015 Elsevier Inc. All rights reserved.

introduced by Skowronski and Harris, who also proposed some modifications to the Mel scale filterbank, which resulted in a significant improvement over the original MFCC. In the attempt to make noise-robust MFCCs, each of the Mel filters was associated with a weight based on the signal-to-noise ratio (SNR) [8]. With the objective of improving the discriminability of the phoneme classes using linear discriminant analysis, the authors in [9] provided optimized filterbanks.

Evolutionary algorithms were used for the optimization of filterbank features for cases of speech verification [10,11], and phoneme classification [12–15] and [16] was proposed by some researchers. In [13], the Mel filterbank was optimized based on the genetic algorithm (GA) to perform a robust classification of noisy speech using HMM. The scale and center frequencies of filterbanks were modified in the aforementioned study. A similar approach was employed in [14], where the amplitude of the Mel filters was optimized using the same strategy. The authors employed splines to reduce the number of effective parameters in the optimization of filters. The discriminability of the extracted features is another effective parameter for developing ASR systems. Based on this idea, [17] proposed discriminative feature extraction to achieve the optimization of the lifter shape, thus improving speech recognition performance.

In another study [15], optimization of triangular filterbank parameters of MFCC was performed based on particle swarm optimization (PSO) and genetic algorithm. The proposed features were used in Hindi speech recognition in clean and noisy environment. The optimized filterbanks outperformed the conventional MFCCs in terms of recognition accuracy. In [16], the authors proposed an adaptive band filterbank for robust speech recognition. For this purpose a bark-scale filterbank was optimized using GA. the front process for the adaptive filter was the zero crossing peak amplitude.

Despite all attempts made by current researchers, an ongoing challenge remains in developing an ASR system with optimal performance in the presence of noise. This challenge can be attributed to the fact that the optimal features for robust speech recognition are yet to be determined. Moreover, the relationship between the classification error and the features is rarely considered in such methods. In this paper, the key features used to obtain optimal class representation are considered relative to the aspects of discriminability and robustness to noise. The first aspect is realized by using different but complementary representation of speech signals, whereas the later aspect is realized by fusing the decisions of the classifiers as a post classification stage.

Another important issue in obtaining speech features is the optimization framework. The use of evolutionary algorithms in this case has been shown to be helpful [13,14]. However, a general weak point of using of such algorithms is that they need considerable computational time. To mitigate this issue, subset selection methods are used to reduce the number of samples processed in each generation. Although such methods degrade the generalization capability of optimization, they save on computational cost. As a result, a compromise between generalization capability and computational cost is unavoidable in cases of huge database optimization [13].

Despite the achieved feature optimization improvements in evolutionary optimization frameworks [13], the acquisition of the best parameters for the feature extraction method cannot be guaranteed. Additionally, classifier training, which serves an important function as an optimizing objective function, is another aspect to be considered in such frameworks. In this study, these issues are circumvented through a hierarchical evolutionary optimization strategy, in which the evolutionary optimization of the filterbank is performed to identify the minimum classification error for all phonetic classes. The optimal solution is an individual chromosome that encodes the optimizing parameters, thus resulting in a classification error for the phonetic classes. A single optimized filterbank may not discriminate among all classes. Thus, the next step is to find a complementary filterbank for the first optimal solution. Thereafter, another filterbank complementary to the first two is identified. After obtaining a few filterbank parameters, each one is employed by an individual expert to perform classifications as ensembles of other experts. Consequently, a fusion of decisions is required to provide an overall decision for a given speech sample.

Fusion of information has been proposed frequently in literature. For example, Benediktsson and Kanellopoulos [18] introduced a multisource classifier based on a combination of a number of statistical classifiers. In this method, two preliminary classifiers trained with different sources are used to assess the membership of testing samples. In case of agreement of the classifiers on the evaluated class, their decision is accepted; otherwise, a post- classifier is employed to make the final decision. A method for combining multiple sources based on their classification accuracies has been proposed by Lisini et al. [19]. In this context, some methods proposed utilizing fuzzy aggregation rules as well as fuzzy set theory and fuzzy fusion to deal with the uncertainty of the classifier's output [20,21].

The purpose of these techniques is to extract information from various data sources and feature spaces, which is unachievable from single data sources, by using different decision methods or multiple classifiers [22].

In some approaches, decision fusion is performed using consensus theory, wherein the estimations provided by different data sources is aggregated by a single probability function [23]. Linear and logarithmic consensus pools are frequent consensus rules that utilize the weighted sum and the products of the sources, respectively. Therefore, the selection of the weights is an important issue in these methods and involves the measurement of the goodness of the processing data [24].

For phoneme recognition based on speech data, we employed a fuzzy set theory and fuzzy data fusion to aggregate the decisions made by a few classifiers (experts). Before performing the fusion, each expert has been optimized to extract complementary information for decision making on the phonetic class of a test sample. The decision fusion framework adapted for this study employs fuzzy logic to address the uncertainty of each classifier. Thus, the framework can handle conflicting sources of information and make an overall decision based on the opinion of each expert.

The remainder of this paper is organized as follows: First, some basic concepts regarding evolutionary algorithms, particularly the GA, are explained. In the subsequent section, the proposed method is discussed in detail. The results and discussions are presented in Section 3, followed by the conclusion in Section 4.

2. Material and methods

2.1. Speech corpus

In this study, phonetic samples from the TIMIT speech database [25] were used. Speech samples from all dialect areas were randomly selected from this database. However, the selection should include a sufficient number of samples from both male and female speakers. An advantage of using the TIMIT database is that all the speech data have been manually transcribed with extreme care. For the selection of noisy samples, different types of noise with different SNR values were added to the speech samples. A Hamming window was used to extract the frames containing 400 samples with a 25 ms frame length and 10 ms step length. The sampling frequency was 16 kHz. The enhancement of the discriminative capability of the experts is a major objective of this work. Thus, some of the phonemes employed for this study were selected