Digital Signal Processing 31 (2014) 102–114

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Fuzzy-based discriminative feature representation for children's speech recognition

Seyed Mostafa Mirhassani [*], Hua-Nong Ting

Biomedical Engineering Department, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Available online 9 May 2014

Keywords:
Children's speech recognition
Mel-frequency cepstral coefficient
Multi-layer perceptron
Hidden Markov model
Fuzzy-based feature selection

ABSTRACT

Automatic recognition of the speech of children is a challenging topic in computer-based speech recognition systems. Conventional feature extraction method namely Mel-frequency cepstral coefficient (MFCC) is not efficient for children's speech recognition. This paper proposes a novel fuzzy-based discriminative feature representation to address the recognition of Malay vowels uttered by children. Considering the age-dependent variational acoustical speech parameters, performance of the automatic speech recognition (ASR) systems degrades in recognition of children's speech. To solve this problem, this study addresses representation of relevant and discriminative features for children's speech recognition. The addressed methods include extraction of MFCC with narrower filter bank followed by a fuzzy-based feature selection method. The proposed feature selection provides relevant, discriminative, and complementary features. For this purpose, conflicting objective functions for measuring the goodness of the features have to be fulfilled. To this end, fuzzy formulation of the problem and fuzzy aggregation of the objectives are used to address uncertainties involved with the problem.

The proposed method can diminish the dimensionality without compromising the speech recognition rate. To assess the capability of the proposed method, the study analyzed six Malay vowels from the recording of 360 children, ages 7 to 12. Upon extracting the features, two well-known classification methods, namely, MLP and HMM, were employed for the speech recognition task. Optimal parameter adjustment was performed for each classifier to adapt them for the experiments. The experiments were conducted based on a speaker-independent manner. The proposed method performed better than the conventional MFCC and a number of conventional feature selection methods in the children speech recognition task. The fuzzy-based feature selection allowed the flexible selection of the MFCCs with the best discriminative ability to enhance the difference between the vowel classes.

1. Introduction

Industries and regular people are adopting the application of automatic speech recognition (ASR), such as voice-controlled systems for mobility-impaired people, voice dialing, simplified systems based on speech communication, content-based spoken audio search, speech therapy [1], and others. Despite the versatility of the English language, efforts on developing ASR systems are not limited to English. In several languages, such as Chinese [2,3], Japanese [4,5], Thai [6,7], Portuguese [8], and Arabic [9], research in this area continues to develop efficient ASR systems. While most speech recognition systems focus more on adults than children, the speech recognition of children has numerous applications, such as in educational games for children [10–12] aids for pronunciation [13,14], and in reading based on interactive platforms [10,15]. The speech sounds of children have higher spectral variations and more dynamic characteristics than that of adults due to the extension of the vocal tract size in growing children [16,17]. The performance of ASR systems adapted by adult speech degrades under the context of children speech. To deal with this effect in speech recognition, some approaches used a larger number of samples [18]. Variation of acoustic parameters is larger in children compared to that of adults [19,20], and thus, a frequency warping algorithm is used to reduce the variation of acoustic parameters [21]. In this study, a discriminative feature extraction method is proposed to deal with recognition of children's speech.

Conventionally, the extraction of helpful information about speech signals utilizes various signal-processing techniques to investigate relevant signal characteristics such as energy and

[*] Corresponding author. Fax: +60 3 79674579.
E-mail addresses: mostafamirhassani@gmail.com, S.M.Mirhassani@siswa.um.edu.my (S.M. Mirhassani).

http://dx.doi.org/10.1016/j.dsp.2014.05.004
1051-2004/© 2014 Elsevier Inc. All rights reserved.

spectrum. Mel-frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) parameters are the commonly used feature extraction methods. However, the MFCC does not provide an optimal feature space for the purpose of discrimination. Optimization of speech features such as MFCC is studied in some approaches. Occasionally, the extraction and the optimization of the features are simultaneously accomplished [22–25,29]. The reason for developing such structures is that the degradation of ASR systems can be the result of a mismatch between the acoustic conditions of training and application environments including additive noise, channel distortion, different speaker characteristics, and so on. To develop speaker independent ASR algorithms, speaker normalization is accomplished to produce a pitch-independent representation of speech [26–28]. In this way, several speaker normalization approaches utilize the formant-ratio theory. According to this theory, the quality of the spoken vowel depends on the log frequency intervals between formants (defined as a ratio). Consequently, invariant representation of vowels can be realized by shifting activations along a log frequency axis [29–35]. Computing MFCC involves a somewhat spectral smoothing carried out by triangular filter banks on the spectrum of the speeches. Consequently, filters with higher bandwidth result in higher spectral-smoothing and fewer filters as well as fewer coefficients (MFCCs). This effect is responsible for the degradation of the ASR approaches in the context of children's speech. To cope with this effect we employed narrower filters in obtaining MFCCs thus loss of spectral information can be prevented. However, this method results in many redundant and non-informative features.

In various intricate application tasks, such as phoneme recognition where the systems are developed based on real data, processing a large number of features is frequent. However, many of the features are not relevant to the problem of interest. In addition, numerous features demand a significant amount of computations, which slow down the overall process. Under these circumstances, automatically dismissing irrelevant features is necessary to achieve a model that is accurate and reliable in solving the problem at hand [36,37]. Additionally, using the feature selection technique reduces the computational cost, which enhances the response time of the process. Feature selection (FS) is one of the most productive research areas and has attracted a considerable amount of attention over the past three decades. Proposed FS methods in the literature vary in terms of evaluation criteria for the selected subsets. Frequent criteria recommended in the literature include relevance [38], gain entropy [39], and contingency table analysis [40]. These methods offer no intrinsic order of features. For the FS task, two well-known methods for feature evaluation are used. The first one uses distance metrics to measure the overlap between different classes [41]. Under this strategy, probability density functions of the sample distribution can also be considered. Consequently, the subset for which the average overlap is minimal is considered as a solution. Meanwhile, intra- as well as inter-class distances can be measured by considering the fuzziness and entropy of the features [42]. Through the second method, classification errors based on the feature subset candidates are evaluated. Consequently, the subset with minimal misclassification is selected as a solution. However, given that 2^n feature subsets can be generated by n features, we have to conduct a number of computations to obtain the optimal subset. Consequently, numerous related techniques have been proposed in the literature. Several methods in [43] have been compared and a Genetic algorithm (GA) method has been implemented for the variable selection. As recommended in this study, GA is a good choice for large-scale optimization in which the number of variables exceeds 50. In some feature selection approaches, particle swarm optimization [44,45] and ant colony optimization [46–48] have been employed. In different areas of researches multi-objective optimization based on GA have been performed [49] and dealing with conflicting objectives has been studied. Toward the problem of feature subset selection, multiple objectives are defined in several cases. Consequently, compromising between these objectives is necessary to solve the problem. Meanwhile, using flexibilities to define the optimization problem is helpful. For this purpose, fuzzy set theory is used to codify the flexibilities for the objective functions. This technique leads to achieving extra trade-off to solve this problem. Fuzzy optimization techniques have been frequently used in the optimization of conflicting objectives [48,50–52]. In most cases, having a priori knowledge about the priority of the aggregating objectives ensures incorporation of these techniques in an accumulative structure. Thus, combining the multiple objectives in such approaches is unnecessary. An effective idea for weighing the importance of conflicting objectives is using fuzzy decision making. Through fuzzy logic, uncertainties associated with different objective functions can be effectively formulated to combine using fuzzy aggregation functions [53]. In this study, specific fuzzy codification is proposed based on the data structure in the feature space, which considers the statistical dependence of the selecting features. Therefore, complementary discriminative features can be properly selected to enhance the performance of the proposed children's speech recognition method.

This paper is organized as follows. Section 2 provides a brief explanation of the MFCC as the speech feature extraction method. Additionally, a proposed fuzzy-based feature selection method is presented in Section 2. MLP-based speech recognizer as well as HMM-based speech recognizer are explained in Section 3. Experiments are presented in Section 4 and results are discussed in Section 5, while the conclusion is discussed in Section 6.

2. Feature representation

2.1. Speech feature extraction

Providing a rational representation of the speech information, known as speech feature extraction, is the first step to accomplish any recognition procedure. One of the most popular acoustic feature extraction methods widely used in various ASR systems is the MFCC. Although its name seems incomprehensible, it conveys its nonlinear characteristic based on Mel distances. Mel distances or Mel scales are obtained from the resolution of human hearing. The Mel scales have a larger spread from higher frequencies to lower frequencies. In other words, its frequency scale has linear frequency spacing below 1 kHz and logarithmic spacing for more frequencies. Therefore, people can discriminate bass sounds better than treble, and MFCCs can be obtained by a Fourier transformation of the log–log warped frequency spectrum.

The approximate formula to obtain the Mels for a given frequency f in Hz is given as follows:

$$F_{mel} = 2595.\log_{10}\left(1 + \left(\frac{f}{100}\right)\right) \tag{1}$$

To realize the MFCC in the given DFT of the input signal in Eq. (2), the Mel-frequency filterbank is defined with p triangular filters m_j $(j = 1, 2, \ldots, p)$, as shown in Fig. 1.

$$X_a[k] = \sum_{n=0}^{N-1}[n]e^{-j2\pi k/N}, \quad 0 \le k \le N \tag{2}$$

The filter bank is applied to the FFT by multiplying each FFT magnitude coefficient to its corresponding filter gain from the Mel filter bank followed by a summation of the result. This operation is performed by using Eq. (3):