Knowledge-Based Systems 115 (2017) 5–14

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Deep neural network framework and transformed MFCCs for speaker's age and gender classification

Zakariya Qawaqneh[a], Arafat Abu Mallouh[a], Buket D. Barkana[b,*]

[a] Computer Science and Engineering Department, School of Engineering, University of Bridgeport, Bridgeport, CT 06604 United States
[b] Electrical Engineering Department, School of Engineering, University of Bridgeport, Bridgeport, CT 06604 United States

ARTICLE INFO

Article history:
Received 27 April 2016
Revised 28 September 2016
Accepted 7 October 2016
Available online 20 October 2016

Keywords:
Deep neural network
DNN
I-Vector
MFCCs
Speaker age and gender classification

ABSTRACT

Speaker age and gender classification is one of the most challenging problems in speech processing. Although many studies have been carried out focusing on feature extraction and classifier design for improvement, classification accuracies are still not satisfactory. The key issue in identifying speaker's age and gender is to generate robust features and to design an in-depth classifier. Age and gender information is concealed in speaker's speech, which is liable for many factors such as, background noise, speech contents, and phonetic divergences. The success of DNN architecture in many applications motivated this work to propose a new speaker's age and gender classification system that uses BNF extractor together with DNN. This work has two major contributions: Introduction of shared class labels among misclassified classes to regularize the weights in DNN and generation of transformed MFCCs feature set. The proposed system uses HTK to find tied-state triphones for all utterances, which are used as labels for the output layer in the DNNs for the first time in age and gender classification. BNF extractor is used to generate transformed MFCCs features. The performance evaluation of the new features is done by two classifiers, DNN and I-Vector. It is observed that the transformed MFCCs are more effective than the traditional MFCCs in speaker's age and gender classification. By using the transformed MFCCs, the overall classification accuracies are improved by about 13%.

1. Introduction

Currently, computerized systems such as language learning, phone ads, criminal cases, computerized health and educational systems are rapidly spreading and imposing an urgent need for better performance. Such applications can be improved by speakers' age, gender, accent, and emotional state information [1–3]. Age and gender recognition is defined as the extraction of age and gender information from speaker's speech. A key stage in identifying speakers' age and gender is to extract and select effective features that represent the speaker's characteristics uniquely. Another key stage is classifier design. A classifier uses the extracted features to predict the speakers' age and gender.

Numerous feature sets have been developed and evaluated in the literature for this problem. Those features can be classified into three categories, spectral, prosodic, and glottal features. One of the most recognized feature sets is MFCCs which represent the spectral characteristics of speech utterance. MFFCs are widely used in the literature for different speech processing applications such as speech recognition, speaker identification, and noise classification. MFCCs represent the spectrum that is related to vocal tract shape and do not capture the prosodic information [4]. The effectiveness of MFCCs comes from the ability to model the vocal tract in short-time power spectrum. Although previous studies have presented some improvements in this field, the classification of speaker's age and gender has a big room for improvement. More effective feature sets, especially for short-time duration speech utterances, and classifier designs are required to improve current classification accuracies. There are studies reporting high overall classification accuracies [5] (around 90%), however these studies either used a small private corpus or predicted a small number of age and gender classes. AGender database is one of the most challenging databases in speaker's age and gender classification since it

Abbreviations: DNN, Deep neural network; aGender, Age-annotated database of German telephone speech; HTK, Hidden Markov model toolkit; MFCCs, Mel frequency cepstral coefficients; RBM, Restricted Boltzmann machine; DBN, Deep belief networks; GMM, Gaussian mixtures models; SVM, Support vector machines; MLLR, Maximum likelihood linear regression; TPP, Tandem posterior probability; UBM, Universal background model; PPR, Parallel phoneme recognizer; MAP, Maximum-a-posteriori; BNF, Bottle-neck feature; BB-RBM, Bernoulli-Bernoulli RBM; GB-RBM, Gaussian-Bernoulli RBM.

* Corresponding author.
E-mail addresses: zqawaqne@my.bridgeport.edu (Z. Qawaqneh), aabumall@my.bridgeport.edu (A.A. Mallouh), bbarkana@bridgeport.edu (B.D. Barkana).

http://dx.doi.org/10.1016/j.knosys.2016.10.008
0950-7051/

is text-independent; background noise is present; the number of utterances varies for each class; and there are seven classes. The highest reported classification accuracy for this database is around 60% by using a combination of several feature sets [6].

Last few years, DNNs have been used effectively for feature extraction and classification in computer vision [7–9], image processing and classification [8,10], and natural language recognition [11,12]. In 2006, Hinton et al. [13] introduced the RBM for the first time as a keystone for training DBN. Later, Benjio [14] successfully proposed a new way to train DNN by using auto encoders. DNN has a deep architecture that transforms rich input features into strong internal representation [15]. One of the most recent popular techniques is the eigenvoice (I-Vector) which is based on the process of joint factor analysis [16]. Currently, it is considered as one of the state-of-art in the field of speaker recognition and language detection [17,18]. Eigenvoice adaptation is the main procedure to estimate I-Vector which represents a low-dimensional latent factor for each class in a corpus. A test data is scored by a linear strategy that computes the log-likelihood ratio between different classes.

This paper is organized as follows. A brief literature review is provided in Section 2. In Section 3, the methodology of the proposed work is explained. The classifier design is introduced in Section 4. Experimental results and their analysis are presented in Section 5. The conclusion, challenges, and future work follow in Section 6.

2. Literature review

The problem of age and gender classification was studied early in 1950s [19], but the computer-aided systems for deriving the age and gender information from speech have been developed recently [20,21]. Li et al. [22] utilized various acoustic and prosodic methods to improve accuracies by using two or more fusion systems such as GMM base, GMM-SVM mean super vector, GMM-SVM-MLLR super vector, GMM-SVM TPP super vector, and SVM baseline system. Their GMM system used 13-dimensional MFCCs features and their first and second derivatives per frame as input. Cepstral mean subtraction and variance normalization are performed to get zero mean and unit variance on their database. A UBM and MAP techniques [23] are used to model different age and gender classes in a supervised manner for GMM training purpose. Their system achieved an overall accuracy of 43.1%. The other proposed system by Li et al. [22] is the GMM-SVM mean super vector system that is considered as an acoustic-level approach for speaker's age and gender classification. The GMM baseline system is used for extracting features and for training the UBM model. The mean vectors of all the Gaussian components are concatenated to form the GMM super vectors, and then it is modeled by SVMs. One of the advantages of their work is the usage of two-stage frameworks as in [24], which solve the limitation of computer memory required by large database training instead of directly training a multi-class SVM classifier by using all the high-dimensional super vectors. This system achieved a 42.6% overall accuracy for the aGender database. In the GMM-SVM MLLR system, the MLLR adapts the means of the UBM for each utterance to extract the features of the super vector [25]. SVM is used to model the resulted MLLR matrix super vector. Dimension reduction on the MLLR super vector space is done by linear discriminant analysis. It is important to mention that the MLLR matrix contains speaker's specific characteristics and the contents of this transformed matrix are used as feature super vectors for speaker modeling and age and gender recognition. The MLLR achieved an overall accuracy of 36.2%

The GMM-SVM TPP super vector is calculated as probability distribution over all Gaussian components. In this method, the KL-divergence is used to measure the similarity between vectors. The usage of KL-divergence provides discriminative information, which helps getting information about the age and gender of a speaker. In TPP, UBM models are trained independently. Therefore, each UBM component can model some underlying phonetic sounds [26]. The TPP system's overall accuracy is calculated as 37.8%. The SVM baseline system using 450 dimensional acoustic features [22] and several prosodic features, such as F0, F0 envelop, jitter, and shimmer is designed to capture the age and gender information at prosodic level. This system achieved an overall accuracy of 44.6% for the aGender database.

In [22] it is shown that combining these methods will result in low computational cost. Moreover, a score level fusion of different number of systems is used. Each system has its complementary information from other systems. The highest accuracy (52.7%) is attained when the five systems are combined.

Metze et al. [27] studied different techniques for age and gender classification based on telephone applications. They also compared the performance of their system to human listeners. Their first technique, PPR is one of the early systems which were built to deal with automatic sound recognition and language identification problems. The main core of this system is to create a PPR for each class in the age and gender database. They reported that the PPR system performs almost like human listeners with the disadvantage of losing quality and accuracy on short utterances. Their second technique is based on prosodic features. This technique uses several prosodic features jitter, shimmer, statistical information of the harmonics to noise ratio, and many several statistical information of the fundamental frequency. All these features are utilized and analyzed using a system with two layers. The first layer analyzes the features by using three different neural networks. The second layer processes the output information which has already been produced by the first layer by using dynamic Bayesian network. The system based on prosodic features has shown better performance on variation of the utterance duration. Their third technique is the linear prediction analysis which computes a distance between the formants and the signal spectrum based on the linear prediction cover. The Gaussian distributions of the distance were considered to contain useful information about the age and gender of a speaker. This system has failed due to the fact that young and adult speakers have almost the same Gaussian distribution.

This work shares the same goal with previous works in the literature. The previous systems in [22], which are GMM base, GMM-SVM-Mean supervector, GMM-SVM-MLLR supervector, GMM-SVM-TPP supervector, and SVM baseline system, as well as, the previous systems in [27], which are PPR, prosodic feature, and linear prediction analysis systems used a combination of different popular feature sets to classify speakers' age and gender information. Different than the previous works, our proposed work offers a new feature set that is constructed from the MFCCs and a DNN-based classifier that is designed for speaker's age and gender classification. DNN with a bottleneck layer is used to generate bottleneck features from the MFCCs. These features can be considered as a low-dimensional feature set since the bottleneck layer compresses the MFCCs. In addition, a DNN classifier is designed and used instead of combining several classifiers together for a better classification. In [22], it is reported that the highest accuracies are achieved by combining five systems together. On the other hand, our proposed system achieves higher accuracies by using only one classifier.

3. Methodology

In this section, the generation of transformed features and the suggested regularized DNN weights using shared class labels are explained. We propose an approach to transform existing features into more effective features. MFCCs, their first and second derivatives are used as input features for comparison reasons since most