

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/309340345>


Deep Neural Network Framework and Transformed MFCCs for Speaker's Age and Gender Classification

Article in Knowledge-Based Systems · October 2016
DOI: 10.1016/j.knosys.2016.10.008

CITATIONS
6


READS
217

3 authors:




Zakariya Qawaqneh
University of Bridgeport
13 PUBLICATIONS 39 CITATIONS

SEE PROFILE



Arafat Abumallouh
University of Bridgeport
14 PUBLICATIONS 22 CITATIONS


SEE PROFILE




Buket D. Barkana
University of Bridgeport
88 PUBLICATIONS 349 CITATIONS

SEE PROFILE

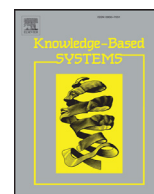
Some of the authors of this publication are also working on these related projects:



Deep Brain Stimulation (DBS) effects on Parkinson 's Disease [View project](#)



Computer-aided diagnosis systems [View project](#)



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%.

© 2016 Elsevier B.V. All rights reserved.

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. MFCCs 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), aaabumall@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/© 2016 Elsevier B.V. All rights reserved.

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

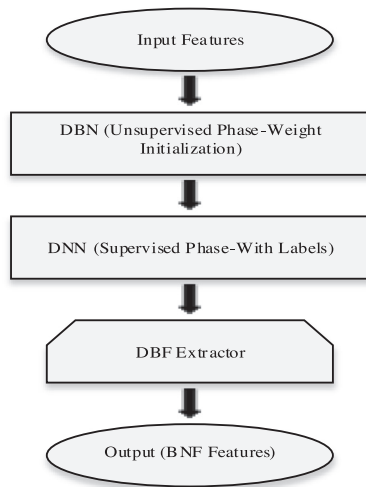


Fig. 1. The main steps for extracting the BNF features from the input features.

of the previous studies have used MFCCs features in age and gender classification [6,22,27].

3.1. Generation of transformed features

New bottleneck features are generated from input features by using DNN as shown in Fig. 1. For example glottal and spectral features can be used to generate a new form of bottleneck features in speech field.

The DNN that is used to generate these features consists of several hidden layers in which one of them has a very small number of units compared to other layers. The resulted features can be considered as a low-dimensional representation since the bottleneck layer compresses the input features and the output labels to form a new bottleneck features. It is as a way of nonlinear dimensionality reduction since it produces a low-dimensional feature set from the input features based on the nonlinear activation functions used to produce the outputs of the units in the neural network. Recently, the usage of bottleneck DNN has shown improved results in auto-encoder to reconstruct the input features [28]. In this paper, the bottleneck features are investigated further and used to classify speaker's age and gender.

In this section, the phoneme label extraction and the BNF extraction are introduced. Firstly, the labels are extracted for each frame for all utterances. Then based on the extracted labels, the

BNF extraction generates the transformed MFCCs using a bottleneck layer in a trained DNN.

3.1.1. Phoneme label extraction (tied-state triphones)

Usually each database has a transcript file for each utterance that contains spoken words. Using the transcript along with speech audio files, the phonemes are extracted and this process is called grapheme-to-phoneme phase. The primary function of the HTK toolkit is to build Hidden Markov Models (HMMs) for speech-based tasks such as recognizers [29]. In the field of speech recognition, the recognition of speech is performed by mapping the sequence of speech vectors to the desired symbols sequence. Several complications may occur while performing the recognition of speech. For example, the mapping between symbols and speech is not one-to-one. In most cases, the speech vector could be mapped to many symbols. Another complication is unclear boundary locations between words in a speech. This will cause incorrect mapping between the speech and the symbols. HTK tool is designed to address such issues using HMMs. HMMs are used to align phonemes with correct labels. It provides word isolation to deal with the unclear boundary location problem. In this work, we utilized the HTK tool in [29] to find the tied-state triphones which will be used later as labels for the output layer in the DNN.

The steps of finding the tied-state triphones is depicted in Fig. 2 and described below.

- Step 1: Generate the monophones by considering all of the pronunciations of each utterance in the database. The pronunciation that matches the best to the speech audio will be selected as an output.
- Step 2: Produce triphones. Monophones are used to produce triphones. The current monophone, X, the previous monophone, L, and the next monophone, R, are processed together.
- Step 3: Generate triphones that do not exist in the training data. These are called tied-state triphones.
- Step 4: Find the best match between each frame of the speech utterance and tied-state triphones. The best match will be the phoneme label of the corresponding target frame.

The phoneme labels are used for speech recognition. In this work, the phoneme labels are used to create transformed features. It keeps the phoneme specific characteristics of each speaker. The phoneme labels also help the DNN to embrace distinctive information in the BNF.

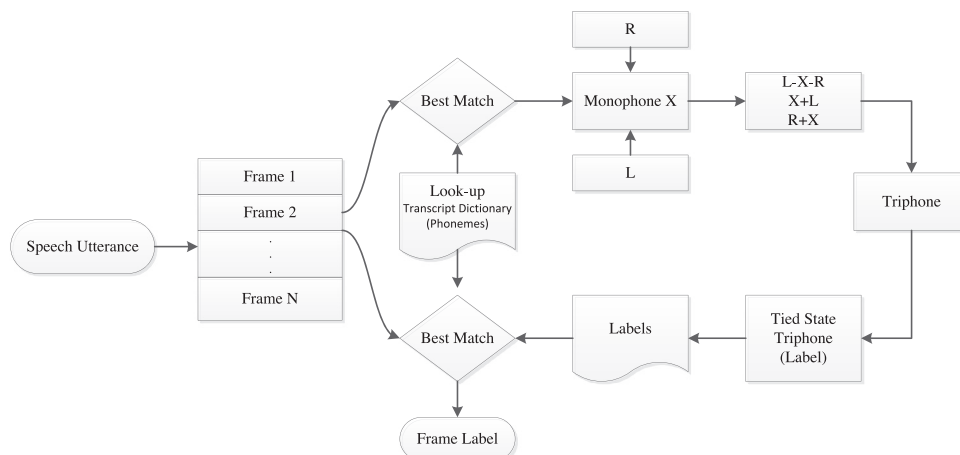


Fig. 2. HTK process for extracting phoneme frame labels.

3.1.2. BNF extraction

In this section, we discuss the BNF extraction process. First we will describe the DNN training procedure in its two phases: the generative (unsupervised) and the supervised. Then, the process of extracting the BNFs features based on the trained DNN will be explained in the BNF extractor section.

A) DNN training

The first phase is generative. The DNN is pre-trained by using an unsupervised learning technique that employs the RBM. The second phase is discriminative. The DNN is trained by using the back-propagation algorithm in a supervised way. An RBM has input layer, V (visible layer) where $V = \{v_1, v_2, \dots, v_V\}$, and the output layer, H (hidden layer) where $h = \{h_1, h_2, \dots, h_H\}$ [30]. The visible and the hidden layers consist of units. Each unit in the visible layer is connected to all units in the hidden layer. The restriction of this architecture is that there is no connection between the units in the same layer. Two types of RBMs, BB-RBM and GB-RBM [31] are used in this work. In the BB-RBM, the visible and hidden layer unit values are binary, $V \in \{0, 1\}$ and $H \in \{0, 1\}$. The energy function of the BB-RBM is defined in Eq. (1)

$$E(v, h) = - \sum_{i=1}^V \sum_{j=1}^H v_i h_j w_{ij} - \sum_{i=1}^V v_i b_i^v - \sum_{j=1}^H h_j b_j^h \quad (1)$$

where V_i is the visible unit in layer i and H_j is the hidden unit in layer j . w_{ij} denotes the weight between the visible unit and the hidden unit. b_i^v and b_j^h are the bias of the visible unit in layer i and the hidden unit in layer j , respectively. For the GB-RBM, the visible unit values are real, where $V \in \mathbb{R}$, and the hidden units values are binary, where $H \in \{0, 1\}$. The energy function of this model is defined as in Eq. (2)

$$E(v, h) = - \sum_{i=1}^V \sum_{j=1}^H \frac{v_i}{\sigma_i} h_j w_{ji} + \sum_{i=1}^V \frac{(v_i - b_i^v)^2}{2\sigma_i^2} - \sum_{j=1}^H h_j b_j^h \quad (2)$$

where σ_i is the standard deviation of the Gaussian noise for the visible unit i . The joint probability distribution which is associated with configuration of (v, h) is defined in Eq. (3)

$$p(v, h; \theta) = \frac{\exp(-E(v, h; \theta))}{Z} \quad (3)$$

θ represent the weights and the biases, while Z is the partition function defined as in Eq. (4).

$$Z = \sum_v \sum_h \exp(-E(v, h; \theta)) \quad (4)$$

The RBM is the basic building block in DBN. It is used as a feature detector and trained in an unsupervised way. The output of a trained RBM is used as an input to train another RBM. Training RBM is very useful for complex problems where the structure of the data is complicated and the implicit features could not be detected directly [32]. A number of RBMs could be stacked together to represent complex structures and to detect implicit features from the previous RBM representation in the stack. The stacked RBMs represent a generative model called DBN. The learning algorithm in the DBN is layer-wise and unsupervised. The layer-wise learning helps to find descriptive features that represent correlation between the input data in each layer [33]. The DBN learning algorithm works to optimize the weights between layers. Moreover, it is proved that initializing the weights between layers in the DBN network enhances the results more than if random weights are used. Another advantage of DBN training lies in its ability to reduce the effect of over-fitting and under-fitting problems where both are common problems in models with big number of parameters and deep architectures. After the DBN learning is completed

and the weights between the layers in the DBN stack are optimized, the supervised training process is started by adding a final layer of labels on top of the DBN layers. These labels represent the final classes of the whole network. In our work, these labels represent the tied-state triphones for the utterance speech data.

B) BNF extractor

BNF architecture is generated from a trained DNN where each layer represents a different internal structure of the input features. In the DNN, the output of each hidden layer produces transformed features. All the layers above the bottleneck layer are removed to produce the BNF extractor as shown in Fig. 3. Fig. 3 explains the proposed bottleneck feature extraction architecture using the phoneme labels. The left side (in Fig. 3) explains the pre-training phase in the DBN consisting of five RBM layers. The first layer is a GB-RBM and the rest are BB-RBM with the bottleneck layer located in the middle. The right side (in Fig. 3) portrays the DNN architecture which is formed by adding a softmax output layer on top of the DBN architecture. The weights for the DNN are tuned during supervised phase.

Introducing bottleneck layer has many benefits as reducing the number of units inside the bottleneck layer, getting rid of redundant values from the input feature set, and reflecting the class labels during the classification process [34,35]. It also helps to capture the descriptive and expressive features of short-time speech utterances [36]. Given a BNF extractor with M layers, the features at the output layer can be extracted using Eq. (5).

$$\begin{cases} l_1(x) = \sigma \left(\sum_{n=1}^N w(x_n + b_1) \right) \\ l_2(x) = \sigma \left(\sum_{n=1}^{F_2} w(x_n + b_2) \right) \\ \vdots \\ l_M(x) = \sigma \left(\sum_{n=1}^{F_M} w(l_{m-1}(x) + b_M) \right) \end{cases} \quad (5)$$

where σ is computed by the logistic function $\sigma(x) = 1/(1 + \exp(-x))$. $X = \{X_1, \dots, X_N\}$ is the feature set vector, and N is the number of input features. L_M is the output of the M th layer. F is a varying number that represents the input for each layer in the BNF extractor. w represents the weights between the input and output nodes in each layer. B represents the bias for each layer.

3.2. Regularizing DNN weights using shared class labels

Traditionally, one label is assigned to each class during the regularization of weights. However in this work, one label is allowed to represent two classes. Those two classes sharing the same label are chosen among the most misclassified classes. By sharing the same label, the weights between the DNN layers are being enforced to converge to an unbiased form with a wider-range representation. Misclassifications between classes are determined by a DNN classifier (Fig. 4A). Two classes having the highest misclassification ratio are chosen to share a label. Let us have a database with seven classes, and the highest misclassifications occurred between classes (3 and 5), and between classes (4 and 6). Therefore, five shared labels are generated, the first label is for the class 1, the second label is for the class 2, the third label is a shared label between the classes 3 and 5, the fourth label is shared between the classes 4 and 6, and finally the fifth label is for the class 7. As shown in Fig. 4B a second DNN structure calculates the regularized weights. These regularized weights are used as initial weights for the third DNN classifier as shown in Fig. 4C.

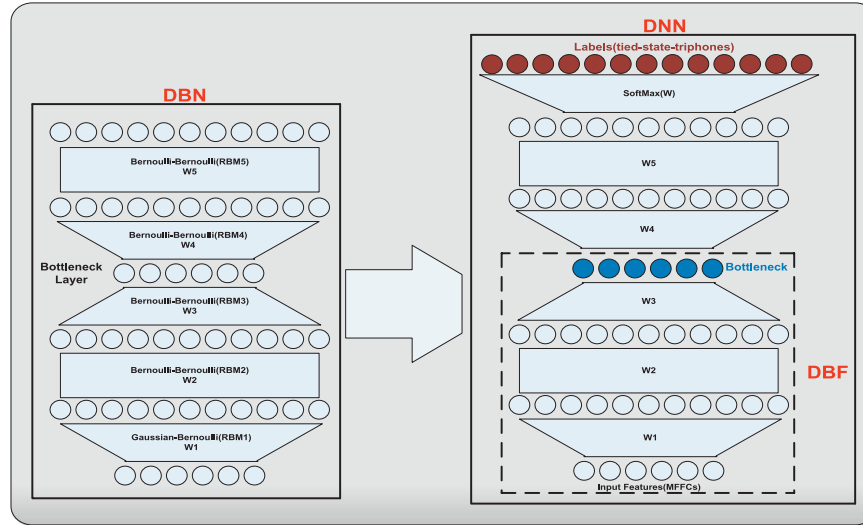


Fig. 3. BNF extractor using trained DNN.

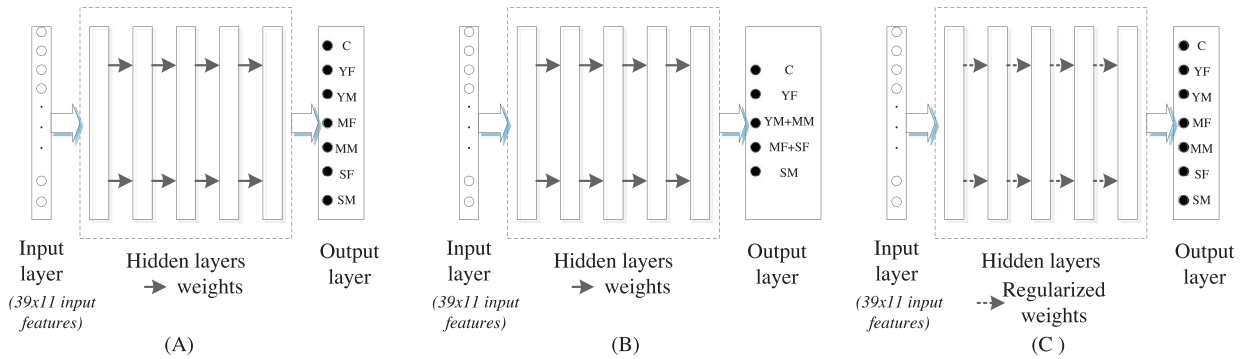


Fig. 4. DNN structures. (A) Finding misclassified classes. (B) Training a second DNN with shared class labels to calculate regularized weights. (C) Initializing a third DNN with regularized weights.

4. Classifier design

I-vector and DNNs are used as classifiers to assess the performance of the transformed MFCCs features. Both classifiers are one of the state-of-art classifiers that have been used in speaker recognition/verification and language identification [10,12,18].

4.1. I-Vector classifier

I-Vector is employed as a back-end classifier in our system. The transformed MFCCs feature set is used as the input vector for the classifier which consists of I-Vectors (eigenvoices) extraction, noise removal, and scoring. I-Vector classifier estimates different classes by using eigenvoice adaptation [37]. The total variability subspace for each utterance is learned from the training data set. Then, the total variability subspace is used to estimate a low-dimensional set from the adapted mean super vectors which are called identity vector (I-Vector). The linear discriminant analysis is applied to reduce the dimension of the extracted I-Vectors by Fisher criterion [38]. For each utterance, GMM mean vectors are calculated. The UBM super vector, M is adapted by stacking the mean vectors of the GMM. It is defined in Eq. (6).

$$M = m + Tw \quad (6)$$

T represents a low-rank matrix, and w represents the required low-dimensional I-Vector. Note that the matrix T is initialized based on the variance of the entire utterances in the training database. After

the extraction of the I-Vectors, noise in each I-Vector is removed by Gaussian probabilistic linear discriminant analysis [39]. Finally, given a test utterance, the score between a target class and the test utterance is calculated using the log-likelihood ratio.

4.2. DNN classifier

Recently, the DNN is considered one of the most popular classifiers and feature extractors. The DNN classifier consists of more than three layers including the input and the output layers. In DNN, each layer is trained based on the features coming from previous layer's output. Therefore, the further the classifier advance in the training and in the layers, the more complex and generative features are generated.

A supervised DNN is built to classify the age and gender for each group on the database based on the frame level. The input feature set for the network is the frames of each class utterances, while the output labels represent the number of the classes in the database. After DNN is trained, all the output activations for each frame for a given class utterance are accumulated and normalized by performing a feedforward process on the trained network to build a model for each class [40]. At the testing process, a new model is created for the utterance test based on the trained network. The cosine similarity between the utterance test model and each class model is computed. Then the final classification decision is made by taking the highest cosine similarity.

Table 1
Age-annotated database of German telephone speech [41].

Class	Age group	Age	Gender
1-C	Children	7–14	Male + Female
2-YF	Young	15–24	Female
3-YM	Young	15–24	Male
4-MF	Middle	25–54	Female
5-MM	Middle	25–54	Male
6-SF	Senior	55–80	Female
7-SM	Senior	55–80	Male

5. Experimental results

In this section, the results of the proposed work are presented. These results are obtained after conducting several experiments on a public database. The database will be discussed in Section 5.1. Section 5.2 explains the feature set in this work. In Section 5.3, the settings for the conducted experiments are discussed in details. Finally, the results of the conducted experiments are presented and discussed in Section 5.4.

5.1. Database

Database of Age and Gender Annotated Telephone Speech, aGender corpus, consists of 47 hours of prompted and free text. The number of speakers in the database is 945 and it includes 7 mixed classes ranging from 7–80 years (Table 1). The number of utterances in the database is 65,364 and the average length of utterances is 2.58 s. The database is divided into two parts; the training/development set contains 53,076 utterances (770 speakers) while the test set contains 17,332 utterances (175 speakers) [41]. The nature of speech content is short commands, single words, and numbers.

5.2. Feature set

MFCCs are widely used in speech signal processing. In literature, most of the speaker's age and gender classification works used MFCCs as input features. For that reason, we chose MFCCs features to evaluate the performance and effectiveness of the proposed approach in generating transformed features. Thus the classification accuracies can be compared with previous findings. Overall classification accuracies of the I-Vector and DNN classifiers are presented in Table 2 by using the traditional MFCCs and the transformed MFCCs feature sets.

5.3. DNN training settings

The utterance is divided into frames of 25 ms. In total, 39 features, one energy and 12- MFCCs with its first and second derivatives, are extracted for each frame. The DNN settings used in this work are based on the work in automatic speech recognition by [34,42]. There are 5 hidden layers with 1024 nodes in each layer except the bottleneck layer where the number of nodes is 39. The number of nodes in the input layer is equal to the length of the

input vector which has $39 \times n$ features. n is set to 11 after rigorous trial and error process. The 11 sequence frames are target frame and the previous and next $(n-1)/2$ frames. The number of nodes in the bottleneck layer is set to the number of input features, which is 39. The number of nodes in the output layer is set to the number of tied-state triphones, which is 4400, in the database. The training data is divided into mini batches. Each mini batch consists of 1024 utterances. 10 epochs are used for training the GB-RBM over all the training data while 12 epochs are used for the rest of the BB-RBMs. The learning rate for GB-RBM and BB-RBM is 0.0025. In the fine tuning phase, 12 epochs are used. The learning rate is initially set to 0.1 for the first 6 epochs, and then it is decreased to one-half its initial value for the remainder epochs.

In this work, the DNN is used as classifier, as well. DNN is problem dependent, therefore, many experiments should be carried out to find the optimal settings for a successful classification. After extensive experiments, DNN is built with 5 hidden layers of 1024 nodes each. The input data is the BNF features, while the number of epochs is 16. The learning rate was initially set to 0.1 for the first 3 epochs, then it is decreased to 0.8 times the old learning rate every two epochs. The momentum value was started at 0.5 for the first 3 epochs and then is increased to 0.9 for the remainder epochs. The same settings are used to find the shared labels between the misclassified classes in order to obtain the initial weights for the classifier as described in Section 3.2.

5.4. Results

Several experiments have been conducted to evaluate the performance of the proposed work. As shown in Table 2, the overall classification accuracy by using the transformed MFCCs is 56.13% and 58.89% by the I-vector and DNN classifiers, respectively. On the other hand, the classification accuracies by using the traditional MFCCs are calculated as 43.60% and 45.89% by the same classifiers. The classification accuracies of MF, MM, SF, and SM classes are increased drastically. The statistical analysis of the MFCCs features is studied by Barkana and Zhou [6] in age and gender problem. They reported that MFCCs features have a near identical distribution and flatness for all age groups of female and male speakers. As a result, the recognition of different age groups becomes difficult by using MFCCs. The transformed MFCCs that are generated for the first time in this work increased the overall classification accuracy by about 13%. One of the reasons for this improvement is that the transformed MFCCs features represent the prosodic features in addition to spectral features. The involvement of the phoneme labels in the generation of the transformed MFCCs made it possible to grasp the prosodic features, such as intonation, stress, tone, and rhythm, of a speaker. Another reason is that the transformed features are the result of using phoneme labels in the training data, and this helped to remove any noise or silent frames so that the transformed features are calculated without acoustic background noise.

Fig. 5 shows the receiver operating characteristics (ROC) of the transformed and traditional MFCCs (with random and regularized weights) by using DNN and I-vector classifiers. The ROC curves

Table 2
The overall classification accuracies of the DNN and I-Vector classifiers using the traditional and the transformed MFCCs (%). Bold values represents the overall performances.

Classifier		C	YF	YM	MF	MM	SF	SM	Overall Acc.
I-vector	Traditional MFCCs	64.86	57.12	49.01	24.50	27.03	49.91	32.80	43.60
	Transformed MFCCs	60.33	66.49	48.00	45.46	48.56	56.89	67.15	56.13
DNN with regularized weights	Traditional MFCCs	54.33	52.60	44.80	25.13	42.33	46.13	55.87	45.89
	Transformed MFCCs	62.23	61.54	53.38	47.69	52.00	64.23	70.77	58.98
DNN with random weights	Traditional MFCCs	56.53	47.27	49.07	27.53	35.33	36.13	53.80	43.67
	Transformed MFCCs	59.69	60.15	48.85	40.08	52.23	60.92	63.38	55.04

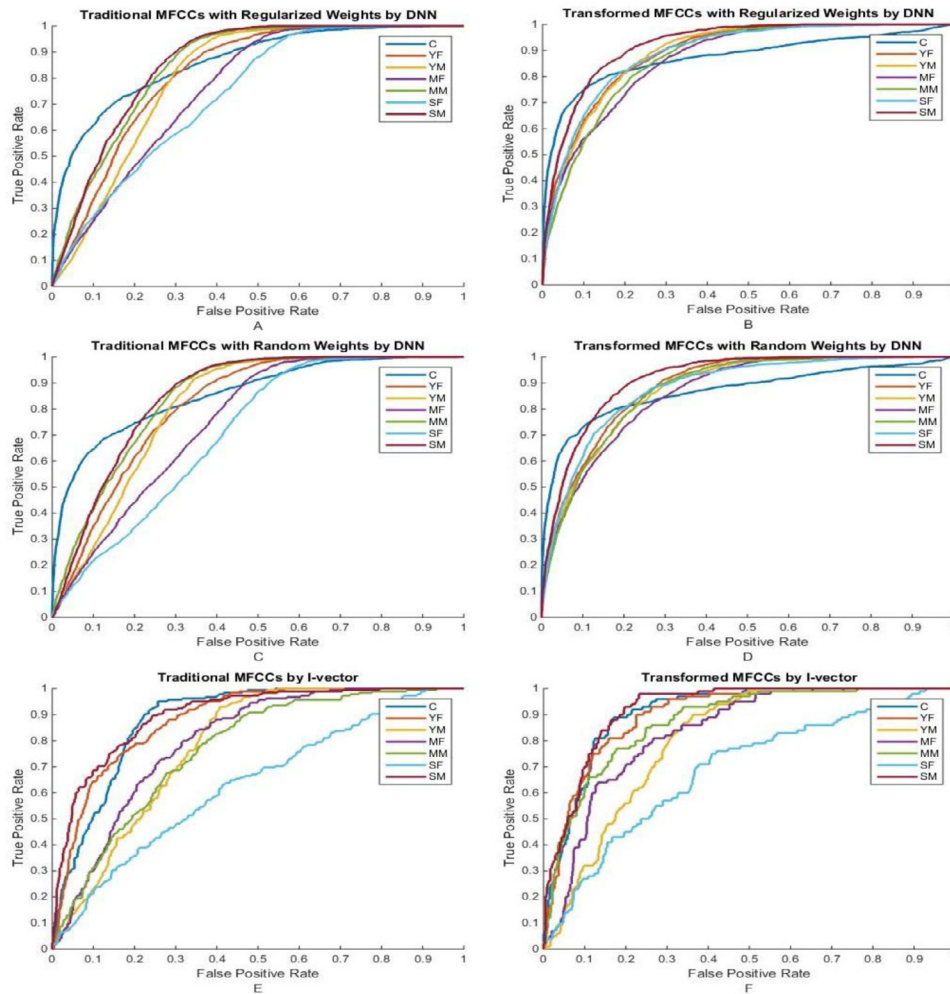


Fig. 5. ROC curves of different classifier scenarios. A) The DNN classifier with regularized weights and the traditional MFCCs. B) The DNN classifier with regularized weights and the transformed MFCCs. C) The DNN classifier with random weights and the traditional MFCCs. D) The DNN classifier with random weights and the transformed MFCCs. E) The I-vector classifier by using traditional MFCCs. F) The I-vector by using the transformed MFCCs.

Table 3

Corresponding AUC measurements for classification of Speaker's age and gender. Bold values represents the overall performances.

Class	DNN regularized weights		DNN random weights		I-vector	
	Traditional MFCCs	Transformed MFCCs	Traditional MFCCs	Transformed MFCCs	Traditional MFCCs	Transformed MFCCs
C	0.86	0.87	0.86	0.87	0.88	0.90
YF	0.81	0.89	0.82	0.88	0.88	0.89
YM	0.81	0.89	0.81	0.87	0.78	0.80
MF	0.76	0.87	0.75	0.86	0.79	0.83
MM	0.85	0.87	0.85	0.87	0.76	0.87
SF	0.74	0.89	0.71	0.88	0.63	0.68
SM	0.86	0.92	0.85	0.91	0.89	0.92
Overall	0.81	0.89	0.80	0.88	0.80	0.84
AUC						

are calculated by using one-against-all rule. The area under curve (AUC) for the transformed MFCCs is found to be bigger than the traditional MFCCs (Table 3 compares the AUC for both sets). The AUC values are calculated as in [43]. The DNN classifier performs better than the I-vector classifier in terms of AUC.

As comparing the classifiers, the DNN classifier performed slightly better than the I-vector classifier. Fig. 6, shows the variance in weights at each layer in the DNN classifier by using random weights and regularized weights. Higher variance between the weights in each layer is needed to distinguish different classes. As it can be seen in Fig. 6, the variance between the weights using shared labels is higher than that of the randomly initialized

weights, therefore, the regularized weights converge faster than the random weights for most of the DNN layers.

Table 4 presents the confusion matrix by using the I-vector classifier. It can be seen that children (C), young female (YF), and senior (SM, SF) classes are classified with higher accuracies compared to the other classes. The major classifications occurred among the same-gender classes. Young female (YF) and senior male (SM) classes have the highest accuracy rates and are correctly classified as 66.49% and 67.15%, respectively. Middle and senior female groups (MF, SF) are classified with the accuracy of 45.46% and 56.89%. Children (C) and young male (YM) classes achieved the accuracy of 60.33% and 48%.

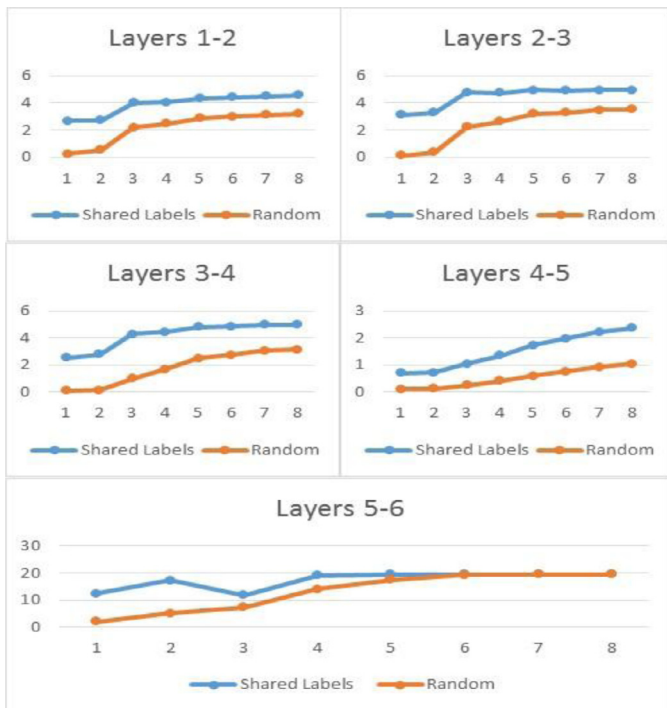


Fig. 6. Variance versus epoch number graphs of regularized and random weights between layers. The x-axis represents the epoch number (1–8), and y-axis represents the variance (y is scaled by 1000).

Table 4

Confusion matrix of the I-vector classifier using the transform MFCCs set (%). Bold values represents the classification accuracies.

Actual	Predicted						
	C	YF	YM	MF	MM	SF	SM
C	60.33	27.90	1.5	4.80	0	2.88	2.59
YF	21.08	66.49	2.70	6.85	0	2.88	0
YM	8.89	1.62	48	0.18	18.97	10.99	11.35
MF	3.60	16.85	2.52	45.46	2.16	29.23	0.18
MM	3.42	1.26	24.43	4.14	48.56	2.34	15.85
SF	7.41	11.17	5.23	13.18	1.44	56.89	4.68
SM	4.50	0.72	11.53	0	15.56	0.54	67.15

Table 5

Confusion matrix of the DNN classifier using the transform MFCCs set (%). Bold values represents the classification accuracies.

Actual	Predicted						
	C	YF	YM	MF	MM	SF	SM
C	63.23	15.38	4.08	3.31	5.08	4.54	4.38
YF	15.92	61.54	0	11.08	0.54	10.23	0.69
YM	1.62	0.62	53.38	2.38	24.46	2.15	15.38
MF	3.38	16.08	2.15	47.69	0.77	28.85	1.08
MM	0.69	0.92	21.77	0.85	52	2.23	21.54
SF	4.69	8.92	1.77	16.23	0.923	64.23	3.23
SM	0.46	0.31	11.85	0.38	13.69	2.54	70.77

Table 5 and Table 6 present the confusion matrices of the DNN classifier using the transformed and traditional MFCCs with regularized weights. In Table 5, the class SM is classified with the highest accuracy (70.77%), while the classes YF, C, and SF are correctly classified with the accuracy ranges between 61% and 64%. The classification accuracies of the MM and YM classes are calculated as 52% and 53.3%, respectively. The lowest accuracy was achieved by the class of MF, as 47.69%. It is observed that the highest misclassification rates have always occurred between the classes with the

Table 6

Confusion matrix of the DNN classifier using the traditional MFCCs set (%). Bold values represents the classification accuracies.

Actual	Predicted						
	C	YF	YM	MF	MM	SF	SM
C	54.33	22.88	2.67	6.13	0.73	11.13	2.13
YF	13.00	52.60	0.40	16.47	0.20	16.93	0.40
YM	0.87	1.00	44.80	2.13	26.20	4.60	20.40
MF	4.40	26.47	1.73	26.13	1.53	37.67	2.07
MM	1.07	0.80	30.93	1.40	42.33	2.60	20.87
SF	4.27	16.20	3.00	23.27	1.20	46.13	5.93
SM	1.07	0.47	10.98	0.67	26.87	4.07	55.87

Table 7

Overall performance comparison in speaker's age and gender classification. Bold values represents the performances of the proposed systems by this work.

System	Overall Acc. (%)
GMM base	43.1
Mean Super Vector	42.6
MLLR Super Vector	36.2
TPP Super Vector	37.8
SVM Base	44.6
MFuse 1 + 2	45.2
MFuse 3 + 4	40.3
MFuse 1 + 2 + 3 + 4	50.4
MFuse 1 + 2 + 3 + 4 + 5	52.7
BNF-I-vector (This work)	56.13
BNF-DNN (This work)	58.98

same gender and close age, or between the children and young female class.

By comparing the classification accuracies of each class in Table 5 and Table 6, the transformed MFCCs help to improve the DNN performance about 10% higher for the classes C, YF, YM, and MM and between 15–20% higher for the classes MF, SF, and SM. This observation can also be seen in the AUC measurements in the Table 3. In their work, Barkana and Zhou [6] reported that traditional MFCCs of the middle-aged female (MF) speakers and senior female speaker have very similar characteristics leading to misclassifications between these two classes. The proposed transformed MFCCs decreased the misclassifications between these two classes significantly since phoneme labels are used in generating the transformed features. The transformed features contain phoneme specific characteristics of each speaker in addition to the spectral characteristics.

Richardson et al. stated in their work [44] that features aligned with phonetical labels or posteriors are still contain speaker-dependent and phonetically discriminative information, which is that are useful for speaker verification. Sarkar et al in [45] reported that “... The results show that the phonetically discriminative MLP features retain speaker-specific information which is complementary to the short-term cepstral features...” Braun et al. [46] studied the effects of the language on estimating speaker's age. They found that the estimation of the speaker's age was language independent, and the listeners did not gain from their knowledge of the corresponding language. We can conclude that the BNFs which are based on phonetical labels retain speaker-dependent information. As a result, BNFs are language independent.

The overall accuracies of the previous studies using the aGender database and the MFCCs feature set are listed in Table 7. The classification accuracies for these systems are reported in Li et al. [22]. The highest reported classification accuracy in the literature is 52.7%, which is achieved by the MFuse 1 + 2 + 3 + 4 + 5 classifier, a combination of GMM base, Mean Supervector, MLLR, TPP, and SVM base systems. GMM baseline, SVM baseline, and GMM-SVM mean

supervector systems have achieved better accuracies than that of the more complex GMM-SVM MLLR supervector system and GMM-SVM TPP supervector system. Combining more systems together did not provide higher classification accuracies. In GMM base system, 39-dimensional MFCCs feature set per frame is extracted. An UBM along with MAP is used to build the class model in a supervised manner. The overall accuracy for that model is reported as 43.1%. The next system, the Mean Super Vector used the GMM baseline system to extract the feature set and training the UBM model. The mean vectors of all the Gaussian components are concatenated to form the GMM supervectors. It is modeled by an SVM. The overall accuracy of this system is stated as 42.6%. In MLLR supervector system, MLLR supervectors and SVM are used to train multi-class models and to score the test set. UBM technique is conducted using the MLLR adaptation for all samples in the training set in order to extract the corresponding MLLR supervectors. The overall accuracy of MLLR system is reported as 36.2% as shown in Table 7. Another variation of the GMM-SVM mean supervector method is the TPP supervector that is calculated as probability distribution over all Gaussian components. In this method the KL-divergence [47] is used to measure the similarity between vectors. In TPP, a UBM model is trained independently as an age and gender models. The overall accuracy for this system is given as 37.8%. In SVM-base system 450 dimensional acoustic features such as F0, jitter, shimmer, along with MFCCs feature set per utterance are extracted. The corresponding features are used as inputs for an SVM classifier by achieving an overall classification accuracy of 44.6%.

The proposed work achieved higher overall accuracies by both BNF-I-vector and BNF-DNN classifiers (56.13%, 58.98%) compared to previous works for the aGender database in the literature. The transformed MFCCs set is proved to be more effective than the traditional MFCCs features in speaker's age and gender classification. There are two main reasons behind this improvement. First, introducing phoneme labels to create BNFs for age and gender problem has a significant impact on the BNFs, which become more discriminative and descriptive. By phoneme labels, phonetic components in a speaker's speech signal have been captured and used in detecting the speaker's age and gender information. Second, the regularized weights converged faster and provided higher variance between classes. These improvements boosted the performance of the classifiers.

6. Conclusions, challenges, and future work

The goal of this paper is to improve the classification accuracies in speaker's age and gender classification. For this purpose, major contributions are made to the area of feature extraction and classifier design. First, a novel approach is introduced to generate transformed MFCCs feature set. Second, classifier weights are regularized by using shared labels. As one of the most popular feature sets in the speech signal processing, MFCCs are proved to be ineffective in speaker's age and gender classification in literature. To improve the performance of the traditional MFCCs, the transformed MFCCs feature set is generated by using BNF extractor. In the BNF extractor, phoneme labels are used to capture phonetic components in the speech. We showed that the DNN can be designed and trained to adapt smoothly with the BNF extractor, so that a new transformed features can be obtained. The shared labels are used to regularize weights between DNN layers. The regularized weights provided faster convergence and higher variance between classes. The performance of the transformed MFCCs is evaluated by two classifiers: DNN and I-Vector. The results showed a significant improvement in the classification accuracies. The overall accuracy of the proposed work is 58.98% and 56.13% for the DNN and I-vector, respectively. To the best of our knowledge, our work is the first

step to apply DNN techniques and I-vector classifier on speaker's age and gender classification problem.

Many challenges were encountered during the development of the proposed work. Tied-state triphones were needed to be used as labels on the DNN output layer. As a solution, a trained GMM-HMM model was used to generate the required labels. The transcript file for each utterance was also required in label generating process. It is noticed that some of the utterance transcripts in the aGender database were incomplete or inaccurate. To address this issue, the database transcript files have been fixed and refined. The optimization of the DNN parameters such as the number of layers, number of units in each layer, weight initialization, learning rate is problem-dependent. The settings of the optimal parameters differ from one problem to another. Therefore, different experiments were conducted to find the optimal settings for the each DNN. Optimizing one parameter alone does not optimize the rest of the parameters. Thus, the parameters should be tuned together to reach the optimal settings. The computation time for training DNN depends on different factors: the size of the database (for speech utterances, there were millions of concatenated frames), the number of features for each sample, number of layers, and number of epochs. To overcome the limited computation resources, we have utilized two of NVIDIA TITAN X GPGPU devices, connected to one single host to make the computation faster by the parallel power of the GPGPUs.

One possible challenge in this work is the extraction of the tied-state triphones. Most of the age and gender speech databases do not come with tied-state triphones. As a result, the tied-state triphones should be carefully extracted to implement the proposed work on a database. The extraction process has to satisfy several requirements such as the manuscripts of the speech utterances and special speech software (such as HTK). The extraction process involves many steps that take time. Another possible challenge of this work is to find the most misclassified classes with each other in order to use the shared labels technique.

The proposed work achieved higher classification accuracies than the previous systems in the literature. This work proves that the bottleneck features can offer better speaker dependent information. In this work, we only utilized MFCCs to generate the bottleneck features. As future work, other time-domain and frequency-domain based features such as fundamental frequency, pitch-range-based, and linear predictive coefficients can be used as input features. It is expected that the classification accuracies would be improved further. In addition, different type of deep neural networks such as convolutional neural networks (CNNs) can be used alone or along with the DNN classifier. In this work, the I-vector was used as a classifier since it is one of the state-of-art techniques in several fields such as language identification and speaker verification. We plan to investigate the usage of the extracted BNF features as input for the I-vector in order to model the corresponding I-vector for each utterance. The resulted I-vectors will be used as a new feature set that could be fed to any classifier. Since the I-vector utilizes techniques such as a within class covariance, it might help to represent the speech utterances in a more distinguishable way. As we plan to fine-tune several DNN architectures jointly by using a new cost function, each DNN will have a different feature set. The DNNs will be trained jointly and simultaneously. It is expected to improve the classification accuracies by representing speech utterances distinctively, since it will utilize more than one feature set for each utterance and include more information about the speaker.

References

- [1] M. Black, A. Katsamanis, C.-C. Lee, A.C. Lammert, B.R. Baucom, A. Christensen, et al., Automatic classification of married couples' behavior using audio features, in: INTERSPEECH, 2010, pp. 2030–2033.

- [2] P. Nguyen, D. Tran, X. Huang, D. Sharma, Automatic speech-based classification of gender, age and accent, in: B.-H. Kang, D. Richards (Eds.), *Knowledge Management and Acquisition for Smart Systems and Services*, vol. 6232, Springer Berlin Heidelberg, 2010, pp. 288–299.
- [3] T. Schultz, Speaker characteristics, in: C. Müller (Ed.), *Speaker Classification I*, vol. 4343, Springer Berlin Heidelberg, 2007, pp. 47–74.
- [4] S.B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *Acoust. Speech Signal Process. IEEE Trans.* vol. 28 (1980) 357–366.
- [5] H.-J. Kim, K. Bae, H.-S. Yoon, Age and gender classification for a home-robot service, in: *Robot and Human interactive Communication*, 2007. RO-MAN 2007. The 16th IEEE International Symposium on, 2007, pp. 122–126.
- [6] B.D. Barkana, J. Zhou, A new pitch-range based feature set for a speaker's age and gender classification, *Appl. Acoust.* 98 (2015) 52–61.
- [7] M.D. Zeiler, *Hierarchical Convolutional Deep Learning in Computer Vision*, New York University, 2013.
- [8] M. Ranzato, G.E. Hinton, Modeling pixel means and covariances using factorized third-order Boltzmann machines, in: *Computer Vision and Pattern Recognition (CVPR)*, 2010 IEEE Conference on, 2010, pp. 2551–2558.
- [9] C. Ekanadham, S. Reader, H. Lee, Sparse deep belief net models for visual area V2, *Adv. Neural Inf. Process. Syst.* 20 (2008).
- [10] G.E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, *Audio Speech Lang. Process. IEEE Trans.* 20 (2012) 30–42.
- [11] T. Deselaers, S. Hasan, O. Bender, H. Ney, A deep learning approach to machine transliteration, in: *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 2009, pp. 233–241.
- [12] D. Yu, S. Wang, Z. Karam, L. Deng, Language recognition using deep-structured conditional random fields, in: *Acoustics Speech and Signal Processing (ICASSP)*, 2010 IEEE International Conference on, 2010, pp. 5030–5033.
- [13] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, *Neural Comput.* 18 (2006) 1527–1554.
- [14] Y. Bengio, Learning deep architectures for AI, *Found. Trends® Mach. Learn.* 2 (2009) 1–127.
- [15] J.M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, et al., Developments and directions in speech recognition and understanding, Part 1 [DSP Education], *Signal Process. Mag. IEEE* 26 (2009) 75–80.
- [16] D. Matrouf, N. Scheffer, B.G. Fauve, J.-F. Bonastre, A straightforward and efficient implementation of the factor analysis model for speaker verification, in: *INTERSPEECH*, 2007, pp. 1242–1245.
- [17] M. Senoussaoui, P. Kenny, N. Dehak, P. Dumouchel, An I-Vector extractor suitable for speaker recognition with both microphone and telephone speech, in: *Odyssey*, 2010, p. 6.
- [18] N. Dehak, P.A. Torres-Carrasquillo, D.A. Reynolds, R. Dehak, Language recognition via i-vectors and dimensionality reduction, in: *INTERSPEECH*, 2011, pp. 857–860.
- [19] E.D. Mysak, Pitch and duration characteristics of older males, *J. Speech Hearing Res.* (1959).
- [20] N. Minematsu, M. Sekiguchi, K. Hirose, Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers, *Acoustics, Speech, and Signal Processing (ICASSP)*, 2002 IEEE International Conference on, 2002 1-137-1-140.
- [21] C. Muller, F. Wittig, J. Baus, Exploiting speech for recognizing elderly users to respond to their special needs, *Eighth European Conference on Speech Communication and Technology*, 2003.
- [22] M. Li, C.-S. Jung, K.J. Han, Combining five acoustic level modeling methods for automatic Speaker's age and gender recognition, in: *INTERSPEECH*, 2010, pp. 2826–2829.
- [23] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, *Digital Signal Process.* 10 (2000) 19–41.
- [24] M. Li, H. Suo, X. Wu, P. Lu, Y. Yan, Spoken language identification using score vector modeling and support vector machine, in: *INTERSPEECH*, 2007, pp. 350–353.
- [25] A. Stolcke, S.S. Kajarekar, L. Ferrer, E. Shrinberg, Speaker recognition with session variability normalization based on MLLR adaptation transforms, *Audio Speech Lang. Process. IEEE Trans.* 15 (2007) 1987–1998.
- [26] X. Zhang, S. Hongbin, Z. Qingwei, Y. Yonghong, Using a kind of novel phonotactic information for SVM based speaker recognition, *IEICE Trans. Inf. Syst.* 92 (2009) 746–749.
- [27] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, et al., "Comparison of four approaches to age and gender recognition for telephone applications," in: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-1089-IV-1092.
- [28] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, *Science* 313 (2006) 504–507.
- [29] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, et al., in: *The HTK Book*, vol. 3, Cambridge University Engineering Department, 2002, p. 175.
- [30] G. Hinton, Training products of experts by minimizing contrastive divergence, *Neural Comput.* 14 (2002) 1771–1800.
- [31] A.-r. Mohamed, D. Yu, L. Deng, Investigation of full-sequence training of deep belief networks for speech recognition, in: *INTERSPEECH*, 2010, pp. 2846–2849.
- [32] L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, *APSIPA Transactions on Signal and Information Processing* 3 (2015) null-null.
- [33] A. Mohamed, T.N. Sainath, G. Dahl, B. Ramabhadran, G.E. Hinton, M.A. Picheny, Deep belief networks using discriminative features for phone recognition, in: *Acoustics, Speech and Signal Processing (ICASSP)*, 2011 IEEE International Conference on, 2011, pp. 5060–5063.
- [34] Y. Bao, H. Jiang, C. Liu, Y. Hu, L. Dai, Investigation on dimensionality reduction of concatenated features with deep neural network for LVCSR systems, in: *Signal Processing (ICSP)*, 2012 IEEE 11th International Conference on, 2012, pp. 562–566.
- [35] F. Grézl, M. Karafiát, S. Kontár, J. Cernocký, Probabilistic and bottle-neck features for LVCSR of meetings, *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007 IV-757-IV-760.
- [36] F. Grézl, P. Fousek, Optimizing bottle-neck features for LVCSR, in: *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4729–4732.
- [37] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, Joint factor analysis versus eigenchannels in speaker recognition, *Audio Speech Lang. Process. IEEE Trans.* 15 (2007) 1435–1447.
- [38] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, *Audio Speech Lang. Process. IEEE Trans.* 19 (2011) 788–798.
- [39] P. Kenny, Bayesian Speaker Verification with Heavy-Tailed Priors, in: *Odyssey*, 2010, p. 14.
- [40] E. Variani, X. Lei, E. McDermott, I. Lopez Moreno, J. Gonzalez-Dominguez, Deep neural networks for small footprint text-dependent speaker verification, in: *Acoustics, Speech and Signal Processing (ICASSP)*, 2014 IEEE International Conference on, 2014, pp. 4052–4056.
- [41] F. Burkhardt, M. Eckert, W. Johanssen, J. Stegmann, A database of age and gender annotated telephone speech, *LREC*, 2010.
- [42] F. Seide, G. Li, X. Chen, D. Yu, Feature engineering in context-dependent deep neural networks for conversational speech transcription, in: *Automatic Speech Recognition and Understanding (ASRU)*, 2011 IEEE Workshop on, 2011, pp. 24–29.
- [43] D.J. Hand, R.J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, *Mach. Learn.* 45 (2001) 171–186.
- [44] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," 2015, arXiv:1504.00923.
- [45] A.K. Sarkar, C.-T. Do, V.-B. Le, C. Barras, Combination of cepstral and phonetically discriminative features for speaker verification, *IEEE Signal Process. Lett.* 21 (2014) 1040–1044.
- [46] A. Braun, L. Cerrato, Estimating speaker age across languages, in: *Proceedings of ICPhS*, 1999, pp. 1369–1372.
- [47] J.R. Hershey, P. Olsen, Approximating the Kullback Leibler divergence between Gaussian mixture models, *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007 IV-317-IV-320.