Combination of Long-Term and Short-Term Features for Age Identification from Voice

Osman BÜYÜK[1], Mustafa Levent ARSLAN[2,3]
[1]*Department of Electronics and Communications Engineering, Kocaeli University, Kocaeli, Turkey.*
[2]*Department of Electrical and Electronics Engineering, Bogazici University, Istanbul, Turkey.*
[3]*Sestek Speech Enabled Software Technologies Incorporation, Istanbul, Turkey*
osman.buyuk@kocaeli.edu.tr

***Abstract*—In this paper, we propose to use Gaussian mixture model (GMM) supervectors in a feed-forward deep neural network (DNN) for age identification from voice. The GMM is trained with short-term mel-frequency cepstral coefficients (MFCC). The proposed GMM/DNN method is compared with a feed-forward DNN and a recurrent neural network (RNN) in which the MFCC features are directly used. We also make a comparison with the classical GMM and GMM/support vector machine (SVM) methods. Baseline results are obtained with a set of long-term features which are commonly used for age identification in previous studies. A feed-forward DNN and an SVM are trained using the long term features. All the systems are tested using a speech database which consists of 228 female and 156 male speakers. We define three age classes for each gender; young, adult and senior. In the experiments, the proposed GMM/DNN significantly outperforms all the other DNN types. Its performance is only comparable to the GMM/SVM method. On the other hand, experimental results show that age identification performance is significantly improved when the decisions of the short-term and long-term systems are combined together. We obtain approximately 4% absolute improvement with the combination compared to the best standalone system.**

***Index Terms*—feature extraction, Gaussian mixture model, neural networks, speech processing, support vector machines.**

I. Introduction

Human speech not only contains linguistic content of the uttered sentence but also other speaker specific information such as speaker's identity, emotional state, gender and age. We utilize all these information sources to improve our daily communication with others. Until now, significant effort has been made to develop automatic systems to extract and recognize the information in the speech for a more natural human-machine interaction. Age information can be used to enhance the user experience in today's voice enabled automatic systems. It can be used to direct the speakers to a specific branch of an interactive voice response (IVR) scenario which is specifically designed for the speaker's age. It can also be used to advertise a suitable product to the user according to his/her age.

Efforts have been made to automatically detect age category from voice. In general, methods initially proposed for other speech processing applications such as speech and speaker recognition have been adapted to age identification with small modifications. Gaussian mixture models (GMM) have been introduced for speaker verification [1] and become the dominant modeling approach until recently. In GMM-UBM method, first a universal background model (UBM) is trained using a large speech database collected from various different speakers. Then, it is adapted to speaker models using target speaker's speech and an adaptation technique such as maximum-a-posteriori (MAP) [1]. Support vector machines (SVM) have also been successfully used for speaker verification [2]. In [2], GMM means are concatenated to obtain a GMM supervector. These high dimensional supervectors are used in the SVM modeling. The proposed GMM/SVM method combines the generative power of GMM-UBM with the discriminative power of SVM. More recently, joint factor analysis (JFA) [3], i-vector [4-5] and probabilistic linear discriminant analysis (PLDA) [6-7] methods have been proposed for mismatch channel compensation in speaker verification. They achieved the state-of-the-art speaker verification performance. Nowadays, deep neural networks (DNN) have been the most popular approach in many machine learning problems thanks to the introduction of efficient methods to train networks with huge number of parameters and availability of large amount of data to perform a robust estimation of the model parameters [8-9]. DNNs and its variants have resulted in accuracy improvement in speech processing applications such as speech recognition [10], speaker/language recognition [11], speech synthesis [12], emotion recognition [13] and spoof detection [14].

Generally, the first stage of a speech processing application is the feature extraction. Speech features can be mainly divided into two categories; short-term and long-term. Short-term features are extracted from a short segment of speech and preserves local information. Mel-frequency cepstral coefficients (MFCC) [15], line spectral pair frequencies (LSPF) [16] and linear prediction coefficients (LPC) [17] are among the most popular spectral short-term features. Additionally, pitch as a prosodic feature is extracted from a short frame of speech and usually combined with the spectral features to improve the performance [18, 19]. On the other hand, long-term features are extracted using longer segments such as the whole utterance. Long-term features typically include informative statistics of the short-term features. Both features have been successfully used for age identification from voice in previous studies [20-22].

In the Interspeech 2010, the paralinguistic challenge is

This work was supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under the project number 3150312.

Digital Object Identifier 10.4316/AECE.2018.02013

organized for age/gender identification from voice [23]. In the challenge, a set of long-term features is defined for age identification. The set contains 1582 long-term features extracted from MFCC, LSPF and pitch-based features. In [21], SVM and NN based systems are trained with the challenge's long-term features. Additionally, the authors train a NN with the PLP and pitch-based short-term features. A GMM-UBM is also trained with modulation spectrogram features [24]. In the experiments, the best performance is obtained with the NN using PLP + pitch features. The accuracy is significantly improved when the systems are fused with linear logistic regression. In [25], age identification performances of GMM-UBM and GMM/SVM methods are investigated. Both systems use MFCC features. In the study, several experiments are run to find the optimum model parameters. The best performance is obtained with the GMM/SVM using 12 MFCC features, 128 mixtures and a linear kernel. In [22], GMM/SVM is compared to GMM-UBM using MFCC features. In the experiments, GMM/SVM outperforms GMM-UBM. In [26], three baseline subsystems, namely, GMM-UBM, GMM/SVM, and SVM with 450 long-term features are compared with four novel subsystems. The novel subsystems are based on i) SVM modeling of UBM weight probabilities, ii) sparse representation of UBM weight probabilities, iii) SVM modeling of GMM maximum likelihood linear regression (MLLR) matrix, iv) SVM modeling of polynomial expansion coefficients of the syllable level prosodic features. Score level fusion of the seven subsystems is also performed. In the experiments, the fused system outperformed all the other subsystems. In [20], following four approaches are compared; i) a parallel phone recognizer (PPR) based on MFCC ii) a dynamic Bayesian network system using prosodic features iii) a system based on the linear prediction (LP) envelope of a windowed speech signal and iv) a GMM based on MFCC. In the experiments, the best performance is obtained with PPR. The PPR method performs as well as human listeners except short utterances. In [27], i-vectors are used for age classification. In the method, average of i-vectors corresponding to each age class is computed during the training stage. The cosine distance between the test and target age class i-vectors is computed in the test. The proposed method achieved a state-of-the-art performance on the Interspeech 2010 paralinguistic challenge age identification database. In [28], a feed-forward DNN is used as a bottleneck feature extractor for age identification. The bottleneck DNN has 5 hidden layers with 1024 nodes in each layer except the last layer where the number of nodes is reduced to 39. As a result, the bottleneck DNN transforms the input features and extracts the most relevant ones for the classification. Mel-frequency cepstral coefficients (MFCC) are used as inputs to the DNN. The compressed features are fed into an i-vector and a feed-forward DNN based classifiers for age identification. The transformed MFCCs achieved 13% performance improvement over the traditional MFCCs. However, the proposed method requires the use of tied-state tri-phone labels from a speech recognizer.

In this paper, we make an extensive comparison of various classification methods for age classification from voice using both short-term and long-term features. The short-term features consist of MFCC and their first order derivatives. We train GMM, feed-forward DNN and long short-term memory (LSTM) recurrent neural network (RNN) methods using the MFCC features. Additionally, an SVM and a feed-forward DNN are implemented using GMM supervectors obtained from the MFCC. To the best of our knowledge, there is no previous study which uses the GMM supervectors to feed a DNN for age identification from voice. The long-term features consist of the Interspeech 2010 paralinguistic challenge features. We choose this feature set since they are commonly used for age identification in previous studies [20-22]. We also combine the age classification decisions of two short-term and three long-term systems for further performance improvement. All the methods are tested using a speech database which includes 228 female and 156 male speakers. In the experiments, it is observed that the proposed GMM/DNN method significantly outperforms all the other DNN architectures. Among all the methods, the best performances are obtained with the GMM/SVM and GMM/DNN methods. The performance is further improved with the system combination. We achieve approximately 4% absolute improvement with the combination when compared to the best standalone system. As a result, 63.51% overall recognition accuracy is obtained with the combined system. The accuracy reaches 77.5% for female speakers.

The main contributions of this paper can be summarized as follows; i) we propose to use GMM supervectors to train a DNN for age identification from voice and show that the performance of the proposed method is significantly better than three other DNN architectures ii) we make an extensive comparison of three state-of-the-art classification methods (GMM, SVM and DNN) for age classification problem using well-known short-term and long-term features iii) we show that the age identification performance can be improved when the information sources from the short-term and long-term systems are combined together.

The remaining of the paper is organized as follows. In Section 2, we introduce our database. In Section 3, our methodology is presented. The implementation details of the method are given in Section 4. Experimental results are provided in Section 5. The last section is devoted to conclusions and future works.

II. DATABASE

Our age identification speech database consists of recordings from 384 speakers. 228 of the speakers are female and 156 of them are male. The youngest speaker in the database is 15 years old and the eldest speaker is 84 years old. The recordings are taken in a soundproofed room to minimize the background noise. They are recorded in 16 kHz, 16 bits, pulse code modulation (PCM) format. Each speaker in the database is required to read 1200 pre-defined Turkish utterances. The length of the utterances ranges from a letter to couple of words. In our experiments, we did not use the entire database. Instead, only 200 utterances of total 1200, which are longer than 3 seconds, are taken for each speaker. The maximum duration of the recordings is approximately 10 seconds.

We define three age categories for the experiments;

young, adult and senior. Age of young speakers is between 15 and 25. It is between 26 and 40 for adult and over 40 for senior speakers. Speakers in each age category are grouped according to their genders. The number of speakers in each age-gender category is summarized in Table I. The experimental results in Section 5 are provided for each gender, separately. In overall case, results for female and male speakers are accumulated.

TABLE I. NUMBER OF SPEAKERS IN THE AGE IDENTIFICATION SPEECH DATABASE FOR EACH AGE-GENDER CATEGORY.

	Female	Male	Total
Young	57	47	104
Adult	123	93	216
Senior	48	16	64
Total	228	156	384

III. METHODOLOGY

A. Feature extraction

In general, the first stage of a speech processing application is the feature extraction. Speech features can be broadly classified into two categories; short-term and long-term. In the short-term feature extraction, speech signal is divided into short frames in order to preserve local information. Then, a feature vector is extracted for each frame. On the other hand, long-term features are extracted using longer segments such as whole utterance. Both feature types have been successfully used for age identification from voice in previous studies [20-22]. In this study, we compare the performances of GMM, SVM and DNN based systems using the short-term and long-term features. The features are extracted using OpenSmile toolkit [29].

i) Short-term features

We use the well-known MFCC as short-term features. In order to calculate the MFCC, the discrete Fourier transform (DFT) of the speech frame is computed. Magnitude DFT coefficients are multiplied by a triangular filter gain. Each triangular filter covers a specific frequency sub-band. The frequency sub-bands are distributed according to mel-scale. Then, the results for each filter-bank channel are accumulated. Finally, discrete cosine transform (DCT) is applied to the logarithm of the accumulated filter-bank energies. Generally, time derivatives of the static parameters are appended to the feature vector for performance improvement. First (delta) and second (acceleration) order derivatives are commonly used in various speech processing applications. In our experiments, we use 13 MFCC (including logarithm of energy) from 26 mel-frequency filter-bank channels. 13 delta coefficients are appended to the MFCC vector resulting in a 26 dimensional feature vector. The features are extracted using 25 millisecond window length and 10 millisecond skip size.

ii) Long-term features

The long-term set contains features from 34 low-level descriptors (LLD). The LLD include PCM loudness, 15 MFCC, logarithmic power of the first 8 mel-frequency bands, 8 LSPF, smoothed fundamental frequency contour's envelope and the final fundamental frequency candidate's voicing probability. Delta coefficients are also appended to the static LLD. 21 functions are applied to each LLD. As a result, 1428 dimensional (68 LLD x 21 functions) feature vector is obtained. Detailed information about the applied functions can be found in [29].

Moreover, 4 pitch-based LLD together with their first order delta coefficients are used in the long-term features. The pitch-based LLD includes the smoothed fundamental frequency contour, frame-to-frame shimmer, frame-to-frame jitter and the delta frame-to-frame jitter. 19 functions are applied to the pitch-based LLD and 152 features (8 LLD x 19 functions) are obtained. Finally the number of pitch onsets and the input duration are appended to the feature vector. The final resulting long-term feature vector has 1582 dimensions (1428 + 152 + 2) for each utterance.

A moving average filter with length 3 is used to smooth the LLD contours. First order delta coefficients are extracted from the smoothed LLD. All the pitch-based LLD are set to zero for unvoiced regions.

B. Classification methods

GMM, SVM, DNN and LSTM have been widely used in various classification problems. In this sub-section, we briefly describe these classification methods.

i) Gaussian mixture models (GMM)

GMM has been extensively used for many pattern recognition applications. In GMM method, the feature vectors are modeled with a mixture of Gaussian distributions. The probability of observing a feature vector given the GMM is computed as in Equation 1;

$$P(\mathbf{o}_t \mid \lambda) = \sum_{i=1}^{M} w_i N(\mathbf{o}_t, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

where $\mathrm{N}(\mathbf{o}_\mathrm{t}, \boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_\mathrm{i})$ is a Gaussian distribution. In Equation 1, $(\mathrm{w}_\mathrm{i}, \boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_\mathrm{i})$ are the mixture weight, mean vector and covariance matrix of i^{th} GMM, respectively. $\mathbf{o}_\mathrm{t}$ is the t^{th} observation vector. M is the number of mixtures in the model. If the feature vectors are assumed to be independent, the recognition score can be computed as in Equation 2;

$$\Lambda(\mathbf{O}) = \sum_{t=1}^{T} \log P(\mathbf{o}_t \mid \lambda) \tag{2}$$

In Equation 2, T represents the total number of feature vectors.

ii) Support vector machines (SVM)

SVM finds the best hyperplane that separate training observations of one class from another. Given a set of feature-label pairs $(\mathbf{x}_\mathrm{t}, \mathrm{y}_\mathrm{t})$, where $\mathbf{x}_\mathrm{t}$ represents the feature vector and $\mathrm{y}_\mathrm{t} \in (+1, -1)$ represents class labels for a two class classification problem, the SVMs try to find the solution for the optimization problem in Equation 3 [30-31];

$$\min_{w,b,e} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} e_i \tag{3}$$

subject to constraints in Equation 4 and Equation 5;

$$y_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b) \geq 1 - e_i \tag{4}$$

$$e_i \geq 0 \tag{5}$$

In the equations, $C > 0$ is the penalty parameter of the error term. The function, Φ, is used to map the training feature vectors into a higher dimensional space. SVM tries to find the separating hyperplane with the maximal margin in that higher dimensional space. Furthermore, $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$ is called the kernel function. Radial basis function (RBF) is one of the well-known kernels described in Equation 6;

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2) \quad (6)$$

where $\gamma > 0$ is a kernel parameter to be optimized.

iii) Feed-forward deep neural networks (DNN)

DNN has become the most popular modeling approach for many machine learning problems especially when efficient methods have been introduced to train networks with many parameters [10], [32]. Additionally, the drastic increase in processing power of the today's computers accelerated the research in deep learning. DNN achieved very good performance in many fields of machine learning. Therefore, many research institutes turned their focus on this emerging field.

A feed-forward DNN is a multilayer perceptron with two or more hidden layers [9]. The neurons in the consecutive layers are fully connected. The power of DNN comes from its ability to learn the data in a hierarchical order in which higher level features are obtained from the lower level features and the same lower level features contribute to many higher level ones [9].

In feed-forward DNN, the activation value of j^{th} neuron in the l^{th} layer can be computed as in Equation 7;

$$a_j^l = \sigma(\sum_k w_{jk}^l a_k^{l-1} + b_j^l) \quad (7)$$

In the equation, w_{jk}^l is the weight between j^{th} and k^{th} neurons in the l^{th} layer, b_j^l is the bias term for the j^{th} neuron in the l^{th} layer and σ is the activation function. Sigmoid activation is frequently used and defined as in Equation 8;

$$\sigma(x) = \frac{1}{1 - e^{-x}} \quad (8)$$

iv) Recurrent neural networks (RNN)

The sequential information in the speech signal is not utilized in the traditional feed-forward DNN. On the other hand, it is well known that the performance of speech processing applications can be enhanced when the co-articulation information is taken into account [33]. RNN is introduced to overcome that major shortcoming of the feed-forward DNN [34]. RNN perform the same task for every element of a sequence, with the output being dependent on the previous computations. A simple RNN architecture is shown in Figure 1;

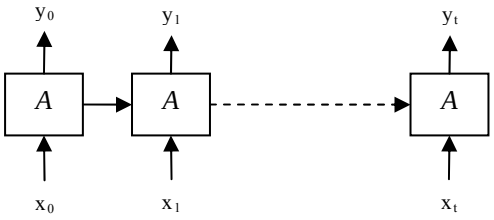

Figure 1. Recurrent neural network (RNN) architecture through time.

In the figure, $\mathbf{x}_t$ and $\mathbf{y}_t$ represents the input and output vectors respectively. The repeating unit, *A*, is shown in detail in Figure 2. In the figure, the small unlabeled rectangular boxes denote either concatenation or copy of the input vector;

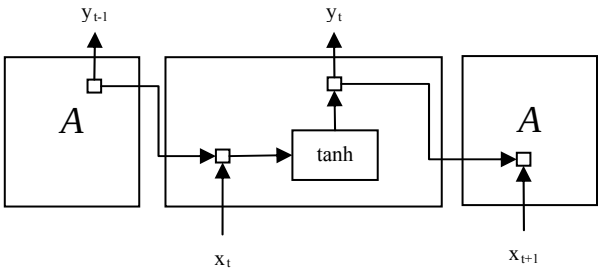

Figure 2. Repeating unit in a recurrent neural network (RNN).

RNN trained with the traditional backpropagation through time algorithm have difficulties learning dependencies between the time steps which are far apart from each other due to what is known as the vanishing gradient problem [35]. Long short-term memory (LSTM) recurrent neural network is proposed to address the vanishing gradient problem [36]. LSTM is a special RNN modified to learn the long-term dependencies. In LSTM, the single neural network layer in the traditional RNN in Figure 2 is replaced with four, interacting layers. The repeating unit in an LSTM is shown in Figure 3 in detail;

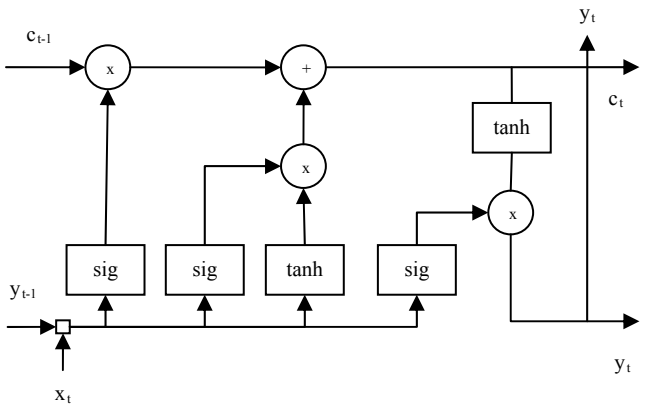

Figure 3. Repeating unit in a long short-term memory (LSTM) recurrent neural network.

The key unit in Figure 3 is the horizontal line running through the top of the diagram. This unit is referred to as cell state. The cell state runs straight down the entire chain, with only some minor linear interactions. The interactions are controlled by the three gates in the figure. The gates are named as forget, input and output from left to right. The

LSTM removes or adds information to the cell state by these gates. The gates are composed of a sigmoid layer and a pointwise multiplication operation. The sigmoid layer outputs a number between one and zero, indicating if the information in the gate will be passed through or not.

IV. Experimental Setup

In this study, a separate modeling is employed for male and female speakers. All age classification accuracies are provided for each gender, separately. We apply a five-fold cross validation to increase the robustness of the obtained results. In the procedure, five separate test sets for each gender is constructed. In each test set, there are two non-overlapping test speakers for each age class. The remaining speakers are used in the model training. An individual train and test are performed for each test set. The test trials in each set are accumulated in order to obtain the recognition results in the following section. In summary, 10 test speakers (2 speakers x 5 test sets) are used in the experiments for each age class. The number of test utterances for each age class is approximately 2000 (2 test speakers x 5 test sets x 200 utterances).

In the following subsections, we summarize the implementation details of classification systems based on short-term and long-term features. In the last subsection, details of the combined system are given.

A. Implementation of the methods using short-term features

We implement two GMM, an SVM and three DNN based systems using the short-term MFCC features.

i) GMM-EM

In the GMM-EM method, an age model is trained for each age category using all training speech samples from the target age class. Expectation maximization (EM) algorithm is used in the training. Initial models are obtained with the Linde-Buzo-Gray (LBG) algorithm [37]. The number of mixtures in GMM is set to 256.

ii) GMM-UBM

In the GMM-UBM method, age models are adapted from an UBM. 18 speakers for each gender are set aside for the UBM training. The UBM speakers are equally distributed among the age categories (6 speakers for each age category). Age models are adapted from the UBM using the remaining speakers' speech samples from the target age class. MAP technique is used for the adaptation. The number of mixtures in GMM is set to 256 as in GMM-EM method. Becars toolkit is used in both GMM implementations [38].

iii) GMM/SVM

In the GMM/SVM method, the UBM in the GMM-UBM is adapted for each training utterance using MAP adaptation. Then, the means of the adapted GMMs are concatenated to obtain 6656 dimensional (256 mixtures x 26 features) GMM supervectors. The GMM supervectors are given as training samples to the SVM. RBF kernel is used in SVM. Kernel parameters are set using a five-fold cross validation. LIBSVM is used for SVM implementation [39].

iv) GMM/DNN

In the GMM/DNN, the same GMM supervectors in the GMM/SVM are fed into a feed-forward DNN. We use a DNN with two hidden layers and sigmoid activation function. The number of neurons in the layers is set to 50 and 100, respectively. These parameters are determined after informal trials.

v) MFCC/DNN

In order to compare with the GMM/DNN method, we trained a feed-forward DNN with raw MFCC features. In the MFCC/DNN method, MFCC features from 11 consecutive frames are concatenated to obtain a 286 dimensional (26 features x 11 frames) feature vector. These feature vectors are directly fed into the DNN. The DNN has two hidden layers with sigmoid activation functions. The number of neurons in the layers is set to 100 after informal trials.

vi) MFCC/LSTM

The same feature vectors in the MFCC/DNN are also fed into a LSTM. The LSTM has two recurrent layers with 100 dimensional outputs. These parameters are selected to keep the LSTM architecture similar to the other DNN architectures. The first recurrent layer has many-to-many and the second layer has many-to-one recurrent model. We applied a dropout of 30% between the recurrent layers in order to avoid the overfitting. The recurrent layers are followed by a fully connected (FC) layer which outputs three classification probabilities corresponding to each age category. We use sigmoid activation in the FC layer.

In MFCC/DNN and MFCC/LSTM implementations, the duration of the speech files are restricted to 3.5 seconds. This is achieved by zero padding the speech feature if the duration is less than 3.5 seconds. If the duration is more than 3.5 second, the speech is cropped. Keras toolkit [40] with Theano [41] is used in all DNN implementations.

B. Implementation of the methods using long-term features

We implement an SVM and a DNN based systems using the long-term features.

i) LONG/SVM

In the LONG/SVM method, 1582 dimensional long-term feature vectors are given as inputs to an SVM classifier. The same kernel function in the GMM/SVM is used in the implementation for a fair comparison. The kernel parameters are optimized with five-fold cross validation.

ii) LONG/DNN

The same long-term features are also fed into a feed-forward DNN. We use the same DNN architecture as in GMM/DNN method for a fair comparison.

C. The combined system

In the combined system, we use the three short-term (GMM-EM, GMM/SVM and GMM/DNN) and the two long-term (LONG/SVM and LONG/DNN) systems. We combine the systems since long-term and short-term features may provide complementary information to each other and thus improve the recognition accuracy. In the combination, we basically apply a weighted summation for the decisions

of the five subsystems. The weights are determined with respect to the recognition accuracy of each subsystem. Final age decision is given according to the highest scoring age class.

V. Results and Discussion

A. Recognition results with short-term features

Percent recognition rates for the short-term MFCC features are provided in Table II for GMM-EM, GMM-UBM, GMM/SVM and GMM/DNN methods. In Table III and IV, confusion matrices in the GMM/SVM method are given for female and male speakers, respectively. In Table III and IV, the rows and columns represent the reference and recognized age classes, respectively. From Table II, it is observed that the recognition performance in male speakers is much lower when compared to female speakers. This is mainly due to the low recognition performance in senior class as observed in Table IV. The recognition performance in senior age class is approximately 20% for male speakers, while it is approximately 52% for female speakers. We think that the main reason for the low performance is the relatively small number of male speakers in senior age category. As given in Table I, there are only 16 male speakers in senior age class. 2 of the speakers are used for testing and 6 of them set aside for UBM training in the GMM/SVM method. Therefore, only 8 speakers are left for the model training. In the future, we plan to extend the database to improve the recognition accuracy. In Table II, it is observed that GMM/SVM and GMM/DNN methods significantly outperform both GMM methods especially for female speakers. This might be due to the fact that both methods combine the powers of the GMM and DNN/SVM.

Table II. Percent Recognition Rates For GMM-EM, GMM-UBM, GMM/SVM and GMM/DNN Methods Using Short-Term MFCC Features.

	Female	Male	Overall
GMM-EM	67.03	49.95	58.49
GMM-UBM	65.00	49.80	57.40
GMM/SVM	75.79	44.03	59.91
GMM/DNN	74.22	44.92	59.57

Table III. Confusion Matrix For GMM/SVM Method Using The Short-Term MFCC Features In Female Speakers.

	Young	Adult	Senior
Young	1712	271	17
Adult	542	1225	233
Senior	31	922	1047

Table IV. Confusion Matrix For GMM/SVM Method Using The Short-Term MFCC Features In Male Speakers.

	Young	Adult	Senior
Young	1369	603	28
Adult	762	1141	72
Senior	406	1186	408

B. Recognition results with long-term features

Percent recognition rates for the long-term features are provided in Table V for LONG/SVM and LONG/DNN methods. In Table VI and VII, confusion matrices in the LONG/SVM method are given for female and male speakers, respectively. As seen from Table V, recognition rates in male speakers are lower compared to the rates in female speaker. In Table VII, it is observed that this is mainly due to low recognition rate in senior male class similar to the short-term feature results. From Table V, we can say that LONG/SVM and LONG/DNN methods perform close to each other. However when we compare Table II and V, we observe that the performance with short-term features is much higher when compared to long-term features especially for female speakers. From the recognition rates we can conclude that the best standalone performances are obtained with the GMM/SVM and GMM/DNN methods using the short-term MFCC features.

Table V. Percent Recognition Rates For LONG/SVM And LONG/DNN Methods Using The Long-Term Features.

	Female	Male	Overall
LONG/SVM	66.40	48.83	57.61
LONG/DNN	65.96	47.28	56.62

Table VI. Confusion Matrix For LONG/SVM Method Using The Long-term Features In Female Speakers.

	Young	Adult	Senior
Young	1663	322	15
Adult	362	1441	197
Senior	7	501	1292

Table VII. Confusion Matrix For LONG/SVM Method Using The Long-Term Features In Male Speakers.

	Young	Adult	Senior
Young	1417	570	13
Adult	844	1031	100
Senior	491	1326	183

C. Comparison of DNN types

In this subsection, we compare the GMM/DNN method with two other DNN architectures, namely MFCC/DNN and MFCC/LSTM. We perform the comparison for only female speakers due to the results in the previous two subsections. The results are provided in Table VIII. As observed in the table, the proposed GMM/DNN significantly outperforms both other DNN architectures in which raw MFCC features are directly used. Additionally, MFCC/LSTM performs slightly worse than MFCC/DNN despite it benefits from long-term dependencies in the speech signal. In the future, we will further analyze the results in the LSTM experiment and try to improve the recognition accuracy with different LSTM network architectures.

Table VIII. Comparison Of DNN Architectures. Percent Recognition Rates Are Provided For GMM/DNN, MFCC/DNN And MFCC/LSTM Methods.

	Female
GMM/DNN	74.22
MFCC/DNN	66.75
MFCC/LSTM	65.50

D. The combined system

Percent recognition rates for the combined system are given in Table IX. As observed in the table, the combined system performs as well as the best standalone systems for both genders. We obtain the best overall recognition accuracy with the combined system. The overall accuracy is 63.51% which is approximately 4% absolute higher than the best standalone system. The overall accuracies with the best standalone GMM/DNN and GMM/SVM systems are

59.91% and 59.57%, respectively. The accuracy reaches 77.50% in female speakers where the amount of data is adequate for each age category.

TABLE IX. PERCENT RECOGNITION RATES FOR THE COMBINED SYSTEM. IN THE COMBINED SYSTEM, AGE CLASSIFICATION DECISIONS OF THE GMM-EM (SHORT-TERM), GMM/SVM (SHORT-TERM), GMM/DNN (SHORT-TERM), LONG/SVM (LONG-TERM) AND LONG/DNN (LONG-TERM) METHODS ARE COMBINED.

	Female	Male	Overall
Combined	77.50	49.52	63.51

VI. CONCLUSION

In this paper, we propose to use GMM supervectors in a feed-forward DNN to classify the age of a speaker from his/her voice. We make an extensive comparison of the method with GMM, SVM, DNN based methods using long-term and short-term features. Short-term features consist of MFCC and their first order derivatives. Long-term features contain statistics of MFCC, LSPF and pitch-based features and commonly used in previous age identification studies. We also combine the short and long term systems for performance improvement. In the experiments, we observed that the best performance is obtained with the GMM/SVM and GMM/DNN methods using the short-term MFCC features. GMM/DNN outperformed all other tested DNN architectures. System combination also improved the performance significantly. Approximately 4% absolute improvement over the best standalone system is obtained with the combination. The recognition accuracy of the combined system reaches 77.5% for female speakers. In the future, we plan to extend our database especially with male speakers. We will also implement an i-vector based age identification system. Moreover, we will investigate features which might be more suitable for age identification task.

REFERENCES

[1] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10 (1–3), pp. 19–41, 2000. doi:10.1006/dspr.1999.0361

[2] W.M. Campbell, D.E. Sturim, D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters, vol. 13 (5), pp. 308-311, 2006. doi:10.1109/LSP.2006.870086

[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel, "A study of inter-speaker variability in speaker verification," IEEE Transactions on Audio Speech and Language Processing, vol. 16 (5), pp. 980–988, 2008. doi:10.1109/TASL.2008.925147

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel , P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Transactions on Audio Speech and Language Processing, vol. 19 (4), pp. 788–798, 2011. doi:10.1109/TASL.2010.2064307

[5] P. Kenny "A small footprint i-vector extractor," in The Speaker and Language Recognition Workshop (ODYSSEY), Singapore, pp. 1–6, 25-28 June 2012.

[6] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in The Speaker and Language Recognition Workshop (ODYSSEY), Brno, Czech Republic, pp. 014, 28 June-1 July 2010.

[7] S.J.D. Prince, J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, pp. 1–8, 14-20 October 2007. doi:0.1109/ICCV.2007.4409052

[8] G.E. Hinton, S. Osindero, Y. The, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, pp. 1527-1554, 2006. doi:10.1162/neco.2006.18.7.1527

[9] L. Deng, D. Yu, "Deep learning methods and applications," Foundations and Trends in Signal Processing, vol. 7 (3-4), pp. 197-387, 2013. doi:10.1561/2000000039

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Processing Magazine, vol. 29 (6), pp. 82-97, 2012. doi:10.1109/MSP.2012.2205597

[11] F. Richardson, D.A. Reynolds, N. Dehak, "Deep neural network approaches to speaker and language recognition," IEEE Signal Processing Letters, vol. 22 (10), pp. 1671-1675, 2012. doi:10.1109/LSP.2015.2420092.

[12] H. Zen, A. Senior, M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7962–7966, 26-31 May 2013. doi:10.1109/ICASSP.2013.6639215

[13] I.J. Tashev, Z.Q. Wang, K. Godin, "Speech emotion recognition based on Gaussian mixture models and deep neural networks", in Information Theory and Applications Workshop (ITA), February 2017. doi:10.1109/ITA.2017.8023477

[14] C. Zhang, C. Yu, J.H.L. Hansen, "An investigation of deep learning frameworks for speaker verification anti-spoofing," IEEE Journal of Selected Topics in Signal Processing, vol. 11 (4), pp. 684-694, 2017. doi:10.1109/JSTSP.2016.2647199

[15] S.B. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28 (4), pp. 357–366, 1980. doi:10.1109/TASSP.1980.1163420

[16] J. Makhoul, "Linear prediction: A tutorial review", Proceeding of the IEEE, vol. 63 (4), pp. 561–580, 1975. doi:10.1109/PROC.1975.9792

[17] F. Itakura, "Minimum prediction residual principle applied to speech recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 23 (1), pp. 67-72, 1975. doi:10.1109/TASSP.1975.1162641

[18] D.A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, B. Xiang, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada, 6-10 April 2003. doi:10.1109/ICASSP.2003.1202760

[19] B. Yegnanarayana, S.R.M. Prasanna, J.M. Zachariah, C.S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," IEEE Transactions on Audio Speech and Language Processing, vol. 13 (4), pp. 575-582, 2005. doi:10.1109/TSA.2005.848892

[20] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J.G. Bauer, B. Little, "Comparison of four approaches to age and gender recognition for telephone applications," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hawaii, Honolulu, USA, 16-20 April 2007. doi:10.1109/ICASSP.2007.367263

[21] H. Meinedo, I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features," in International Conference on Spoken Language Processing (INTERSPEECH), Makuhari, Japan, 26-30 September 2010.

[22] T. Bocklet, A. Maier, J.G. Bauer, F. Burkhardt, E. Noth, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, USA, 31 March-4 April 2008. doi:10.1109/ICASSP.2008.4517932

[23] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mueller, S. Narayanan, "The Interspeech 2010 paralinguistic challenge," in International Conference on Spoken Language Processing (INTERSPEECH), Makuhari, Japan, 26-30 September 2010.

[24] B.E. Kingsbury, N. Morgan, S. Greenberg, "Robust speech recognition using the modulation spectrogram," Speech Comunication, vol. 25, pp. 117-132, 1998. doi:10.1016/S0167-6393(98)00032-6

[25] M. Feld, F. Burkhardt, C. Müller, "Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services," in International Conference on Spoken Language Processing (INTERSPEECH), Makuhari, Japan, 26-30 September 2010.

[26] M. Li, K.J. Han, S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," Computer Speech and Language, vol. 27 (1), pp. 151-167, 2013. doi:10.1016/j.csl.2012.01.008

[27] J. Grzybowska, S. Kacprzak, "Speaker age classification and regression using i-vectors," in International Conference on Spoken

Language Processing (INTERSPEECH), San Francisco, California, USA, 8-12 September 2016. doi:10.21437/Interspeech.2016-1118

[28] Z. Qawaqneh, A.A. Mallouh, B.D. Barkana, "Deep neural network framework and transformed MFCCs for speaker's age and gender classification," Knowledge Based Systems, vol. 115, pp. 5-14, 2017. doi:10.1016/j.knosys.2016.10.008

[29] F. Eyben, M. Wöllmer, B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in ACM International Conference on Multimedia, Firenze, Italy, 25-29 October 2010. doi:10.1145/1873951.1874246

[30] B.E. Boser, I. Guyon, V. Vapnik, "A training algorithm for optimal margin classifiers," in ACM Workshop on Computational Learning Theory, Pittsburgh, USA, pp. 144–152, 27-29 July 1992. doi:10.1145/130385.130401

[31] C. Cortes, V. Vapnik, "Support-vector networks," Machine Learning, vol. 20 (3), pp. 273–297, 1995. doi:10.1007/BF00994018

[32] Y. Bengio, "Learning deep architectures for AI," Foundations and Trends in Machine Learning, vol. 2(1), pp. 1-127, 2009. doi:10.1561/2200000006

[33] O. Buyuk, "Sentence-HMM state-based i-vector/PLDA modelling for improved performance in text dependent single utterance speaker verification," IET Signal Processing, vol. 10 (8), pp. 918-923, 2016. doi:10.1049/iet-spr.2015.0288

[34] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proceedings of the National Academy of Sciences of the USA, vol. 79 (8), pp. 2554–2558, April 1982. doi:10.1073/pnas.79.8.2554

[35] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen Netzen," Diploma thesis 1991, TU Munich.

[36] S. Hochreiter, J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9 (8), pp. 1735-1780, November 1997. doi:10.1162/neco.1997.9.8.1735

[37] Y. Linde, A. Buzo, R. Gray, "An algorithm for vector quantizer design," IEEE Transactions on Communications, vol. 28 (1), pp. 84-95, 1980. doi:10.1109/TCOM.1980.1094577

[38] R. Blouet, C. Mokbel, H. Mokbel, E.S. Soto, G. Chollet, H. Greige, "Becars: a free software for speaker verification," in The Speaker and Language Recognition Workshop (ODYSSEY), Toledo, Spain. pp. 145–148, 31 May – 4 June 2004.

[39] C.C. Chang, C.J. Lin, "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol. 2 (3), pp. 27:1-27, 2011. doi:10.1145/1961189.1961199

[40] F. Chollet, Keras. Github repository 2015. https://github.com/fchollet/keras.

[41] R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky, et. al. "Theano: A Python framework for fast computation of mathematical expressions," arXiv e-prints 2016. 1605.02688: http://arxiv.org/abs/1605.02688.