International Journal of Speech Technology (2020) 23:223–240
https://doi.org/10.1007/s10772-020-09680-4

Multi-task learning DNN to improve gender identification from speech leveraging age information of the speaker

Mousmita Sarma[1] · Kandarpa Kumar Sarma[1] · Nagendra Kumar Goel[2]

Received: 11 August 2019 / Accepted: 27 January 2020 / Published online: 10 February 2020

Abstract
We propose a method which provides age of the speaker as an additional information while training a machine learning model for gender identification. To achieve this objective, we design a multi-task learning Deep Neural Network (DNN) model where the primary output layer has the speakers' gender as target. Further, we use age group of the speaker as auxiliary target for each utterance, where age groups are created considering the gender of the speaker. We experimentally prove that multi-task learning DNN outperforms Gaussian Mixture Model (GMM) or single-task learning DNN trained only for gender recognition for more real life oriented datasets. For such datasets we have recordings of speakers' from all age groups (children to seniors). We use raw speech waveform as input to our DNN which executes the multi-task learning with the freedom to follow gender and age discriminative features during training. The raw waveform front end uses convolutional layer based filter learning. Further, we use Long Short Term Memory cell based recurrent projection (LSTMP) layers for modeling temporal dynamics of speech from learned feature representation.

Keywords Gender identification · Multi-task learning DNN · Age · Raw speech waveform

1 Introduction

Speech signal primarily carries linguistic information in form of the message which is to be conveyed and used for multiple forms of communication and interaction. This information is extracted as a sequence of sound or phone units. However, along with the message, the raw speech signal also carries para-linguistic information such as gender, age, speaker's identity and emotional state. Extraction of such para-linguistic information from speech signal has been accepted to be a critical problem in the domain of Human Machine Interaction (HMI) systems. It has been found that these systems are gaining popularity for the development of emotionally sensitive HMI technology. In the evolving set ups of intelligent commercial dialogue systems and smart call centers, such information can be used as meta data to understand speaker's psychology and response. This in turn will help to make the system adaptable to evolving expectations and necessities of the users.

Gender of the speaker is one of the most important para-linguistic characteristics that is extracted from speech signal. Extensive research have been carried out on gender recognition task since the early 2000s and have been considered as successful (or nearly solved problem) in terms of recognition accuracy compared to other para-linguistic tasks. However, with the development of new modeling approaches and growing challenges in the field of speech based HMI system design, the gender identification task continues to attract more attention time to time (Li et al. 2010, 2013; Meinedo and Trancoso 2010; Kumar et al. 2016; Levitan et al. 2016; Alhussein et al. 2016; Hebbar et al. 2018; Kabil et al. 2018).

This work is focused on the design of end to end multi-task learning Deep Neural Network (DNN) for automatic extraction of gender information from speech. The proposed system leverages speaker's age information to improve gender identification accuracy in case of complex real life oriented datasets which includes speakers from all age groups.

✉ Mousmita Sarma
mousmita.s@gauhati.ac.in

Kandarpa Kumar Sarma
kandarpaks@gauhati.ac.in

Nagendra Kumar Goel
nagendra.goel@govivace.com

[1] Department of Electronics and Communication Engineering, Gauhati University, Guwahati, Assam, India

[2] GoVivace Inc., McLean, VA, USA

During a few preliminary experiments performed separately with single-task learning DNN and Gaussian Mixture Model (GMM), we observe that gender recognition accuracy decreases when the test datasets contains speakers from all age groups, like children, young, adult and seniors covering age range within 4–84 years. However, little observation on human's psychology to recognize gender from speech tells us that human takes the consideration of age of the speaker. For example, in case of kids and growing young male as well as seniors above 80 years, it is difficult for human listeners to discriminate gender only from speech information. Therefore, DNN models which is designed to learn gender of speaker may benefit if guidance on the influence of speaker's age is provided during training. With this motivation, in this work we report the design of a multi-task learning DNN where we use speaker's gender as target in the primary output layer whereas we use age group of the speaker as an auxiliary target for each utterance. Here speaker's age is grouped into 7 different categories considering the gender of the speaker namely children, young female, young male, adult female, adult male, senior female and senior male (ageGender group target hereafter). Thus the DNN is able to learn gender related information using evidences which are supported by both gender and age targets and minimizes total objective loss.

One of the primary challenges in para-linguistic speech processing is the selection of appropriate feature set. The most popular feature set of all time for gender and speaker recognition is the mel-frequency cepstral co-efficients (MFCC). MFCC is created by passing Fourier energies through a mel-filter bank followed by a non-linear log operation and discrete cosine transform. However, the process of creation of mel-filter bank output can be modeled by a Convolutional neural network (CNN) layer either using frequency domain or time domain signal representation as input (Sainath et al. 2013, 2015; Palaz et al. 2015, 2019). Such data driven approaches of filter learning ensures that the learning process to be explicitly based on the objective task to learn (instead of filter response extracted ahead of time). A few works in speech recognition have attempted to use such data driven representation learning from the raw speech signal and achieved equal and sometimes better performances with that of MFCCs (Jaitly and Hinton 2011; Sainath et al. 2013, 2015; Palaz et al. 2015, 2019; Tuske et al. 2014; Golik et al. 2015; Ghahremani et al. 2016). Motivated by such success of feature learning from raw speech data, in this work we use time domain approach of filter learning from raw waveform through CNN. We use this raw waveform front end set up previously for emotion recognition (Sarma et al. 2018) and achieve better accuracy than MFCCs and Fourier log energy. It learns features within the network using a block of convolutional and network-in-network (NIN) layers and jointly optimizes gender identification objective with remaining part of the network. Thus, the multi-task learning process is able to learn evidence of features during training time. Long Short Term Memory cell based recurrent projection (LSTMP) layers are used to model information from long temporal context of speech. However, the work also experiments with time delay neural network (TDNN) for modeling speech temporal dynamics and derive a comparative analysis with LSTMP. On the other hand, we have used two softmax output layers to model the output distribution and are trained to minimize cross-entropy objective loss. Multi-task objective loss is the total loss of these two output layers. Error is back-propagated through all the components and are trained jointly considering raw speech waveform as input. The work also trained DNN models using mel-filter bank's output as feature and derived a comparative analysis with raw waveform based DNN. Further, experimental results are also compared with that obtained from Gaussian Mixture Model (GMM) based gender recognition which uses MFCC features.

The proposed DNN is trained using telephone conversations of the National Institute for Standard in Technology (NIST) 2008 Speaker Recognition Evaluation (SRE) datasets 2008 and a subset of OGI Kids corpus (Shobaki et al. 2000). We evaluate the performance of the proposed solution using the NIST SRE 2010 telephony core wavefiles combined with a subset of wavefiles collected from OGI Kids corpus. This combined SRE 2010 and OGI subset is considered as held out test data for the proposed work which have speakers from age 4 years to 84 years. However, in order to validate the versatility of the models across various datasets, we also derive gender identification evaluation results on NIST SRE 2000 and Switchboard Cellular Part-I (https://catalog.ldc.upenn.edu/LDC2001S13) dataset.

Throughout the work we have studied how interrelation between para-linguistic information like gender and age can be used to improve gender recognition accuracy and design a multi-task learning DNN system to optimize the total objective loss through shared hidden representation. Experimental results prove the efficiency of the proposed method and show how accuracy of gender recognition improves over various age groups while the DNN model learns to discriminate gender with the knowledge of speaker's age.

Rest of the paper is organized as follows. Following here is a description of related works and what motivated us to carry out this work in Sects. 2 and 3. We provide a detailed mathematical model (Sect. 4.1) and algorithm level description (Sect. 4.2) of the proposed multi-task learning DNN solution for gender recognition in Sect. 4. Next, in Sect. 5 we describe the feature learning and temporal modeling process of our DNN set up. Section 6 explains details of the experimental setup. Section 6.1 includes description of datasets. Description of MFCC feature based GMM baseline system is provided in Sect. 6.2. A single-task learning DNN

baseline system which uses Mel-filter bank's output as feature is described in Sect. 6.3. Next, in Sect. 6.4 DNN training considerations are described. Finally, we provide results and analysis of the multi-task learning DNN in Sect. 6.5. In Sect. 7 we summarize the findings of this work and subsequently conclude the description in Sect. 8.

2 Related works

Li et al. (2010), reported results of gender identification as part of the 2010 Interspeech Para-linguistic Challenge using aGender database (Li et al. 2010). It combines five different methods namely GMM based on MFCC features, Support vector machine (SVM) based on GMM mean super vectors, SVM based on GMM maximum likelihood linear regression (MLLR) matrix super vectors, SVM based on GMM Tandem super vectors and SVM baseline system based on the 450-dimensional feature vectors including prosodic features. The work reported 82.38% unweighted accuracy (UA) and 86.27% weighted accuracy (WA) on gender task. Li et al. (2013), improved the gender identification task described in Li et al. (2010) by introducing few additional methods like SVM based on Universal Background Model (UBM) weight posterior probability super vectors using the Bhattacharyya probability product kernel, sparse representation based on UBM weight posterior probability super vectors, SVM based on the polynomial expansion coefficients of the syllable level prosodic feature contours in voiced speech segments (Li et al. 2013). The weighted summation based fusion of these seven subsystems at the score level improves the performance to 85% and 88.4% UA and WA respectively. Another work by Meinedo and Trancoso 2010 presented INESC-ID Spoken Language Systems Laboratory (L2F) Age and Gender classification system (Meinedo and Trancoso 2010) which uses short and long term acoustic and prosodic feature fusions. The system addresses the 2010 Interspeech Para-linguistic Challenge, where gender sub challenge task is to classify speech into three classes child, female and male and achieved 83.1% UA and 86.9% WA. Different classification strategies like GMM-UBM, Multi-Layer Perceptrons (MLP) and SVM etc. were compared. The best results reported were obtained by a calibration and linear logistic regression fusion back-end. The work was later extended as described in Meinedo and Trancoso (2011), where evaluation on I-DASH CA (child abuse) corpus is included (Meinedo and Trancoso 2011). The I-DASH CA corpus is composed of audio extracted from CA domain videos, provided by the Dutch Police forces. They reported a final 28.3% classification error rate on 6 h child abuse (CA) test set. Work by Kumar et al. (2016) explored the usefulness of multi-channel information present in different language channels (English, Spanish, French) in Hollywood movie released in the form of DVD to improve gender classification (Kumar et al. 2016). The work proposed fusion of predictions from various audio channels using Recognition Output Voting Error Reduction (ROVER) method. The work reported 7% absolute improvement in gender classification accuracy using the proposed setup in GMM based gender classifier. Levitan et al. (2016) reported another work on Gender identification, which is focused on using spectral features in conjunction with pitch features to identify gender (Levitan et al. 2016). The work also studied cross-lingual dependency of gender identification and presented a robust gender identification system in German with 93% accuracy. A work by Alhussein et al. (2016) reports gender identification which used information obtained from length of vocal folds to classify gender (Alhussein et al. 2016).

In the recent time, DNN based approaches are observed to be used for modeling of relevant information in speech recognition, emotion recognition etc. In this direction, Hebbar et al. (2018) proposed DNN based gender identification from movie audio data. The work proposed a novel transfer learning strategy in convolutional layer based network (Hebbar et al. 2018). It uses speech activity detection (SAD) knowledge from a bidirectional LSTM recurrent neural network model. Here the SAD labels are obtained by aligning audio with subtitle text of the movie. The work reported 85% WA for movie audio in their best set up. On the other hand, deep feature learning approaches are becoming popular for extraction and classification of information from raw speech signals as mentioned previously. In this direction, Palaz et al. (2015) used CNNs directly trained with the speech signals to estimate phoneme class conditional probabilities and achieves a more robust performance in noisy conditions (Palaz et al. 2019). Ghahremani et al. (2016) used convolutional layer and NIN non linearity based feature extraction from raw speech waveforms for acoustic modeling task (Ghahremani et al. 2016). In the domain of para-linguistics also, Trigeorgis et al. (2016) used raw waveforms for emotion rating in a deep CNN framework (Trigeorgis et al. 2016). Sarma et al. (2018) also proposed time domain raw speech waveform based emotion identification set up (Sarma et al. 2018). In the domain of gender identification, Kabil et al. (2018) produced results on AVspoof and ASVspoof 2015 datasets, using CNN based feature learning in a DNN (Kabil et al. 2018). It has been reported that the system outperforms the acoustics feature based MLP.

3 Motivation and contribution

In this work we identify one new issue in gender identification. Since para-linguistic tasks are related to speaker characteristics (Goel et al. 2018), they are inherently interrelated through speech source and system. One task may provide

evidence of relevant or non relevant features for another task. Therefore, age information of speaker should help to improve the gender identification task. A few initial experiments performed by us also reveals that gender identification performance degrades in case of test dataset which have wide range of speakers from all ages (children to seniors). Even human are not very efficient to recognize gender from children, growing male and senior person's speech. Interspeech 2010 para-linguistic challenge also included children as a separate class from female and male for gender sub challenge. To address this issue of age dependency of gender information, in this work we propose a multi-task learning DNN which uses age as an auxiliary target in a different output layer, where primary output layer optimizes gender identification objective. Learning with age auxiliary task, the proposed DNN generalizes better while dealing with the gender recognition task which is the primary objective.

As described in Sect. 2, for gender identification all previously reported works used cepstral, prosodic and glottal source based features in various architectures. Most of the previous works are focused on deriving new acoustic and high level feature sets to discriminate gender from speech and its fusion at system level. Further, all previous works used non DNN based system to classify gender (except the two recent works by Hebbar et al. 2018; Kabil et al. 2018). Therefore, developing fully end to end DNN based system for gender identification task is still an open challenge. In this work, we have configured gender identification DNN through a data driven approach of feature learning. Time domain raw waveform front end layers (more detail included in Sect. 5) consisting of convolutional and NIN layers are used to learn filters within the network and derive task specific feature representation. When it comes to the design of multi-tasking DNN, it makes more sense to allow the network to learn feature from raw data and optimize multi-task objectives jointly from learned hidden representations.

4 Multi-task learning in DNN

Multi-task learning in DNN has recently become popular where objectives of related tasks are optimized using the same hidden representation (Ruder https://arxiv.org/pdf/1706.05098.pdf). Sharing representations between related tasks, enables the model to generalize better while dealing with one primary task. This is because both the tasks provide evidence of a relevant and irrelevant feature for a particular task. Thus the model gets a better guidance while learning. Similarly, multi-task learning enables joint learning of two related tasks more appropriately, since the feature which one task interprets in a complex way may be interpreted easily by the other. However, choosing appropriate related task is a crucial point to achieve effective learning in multi-task model. In this work, we use speaker's ageGender group as an auxiliary task by introducing a separate output layer in a gender identification DNN model, where primary output layer has speaker's gender as target. The layout of the DNN is shown in Fig. 1. It has three distinct parts:

1. Time domain raw waveform front end layers (front-end block hereafter) as shown in Fig. 2.

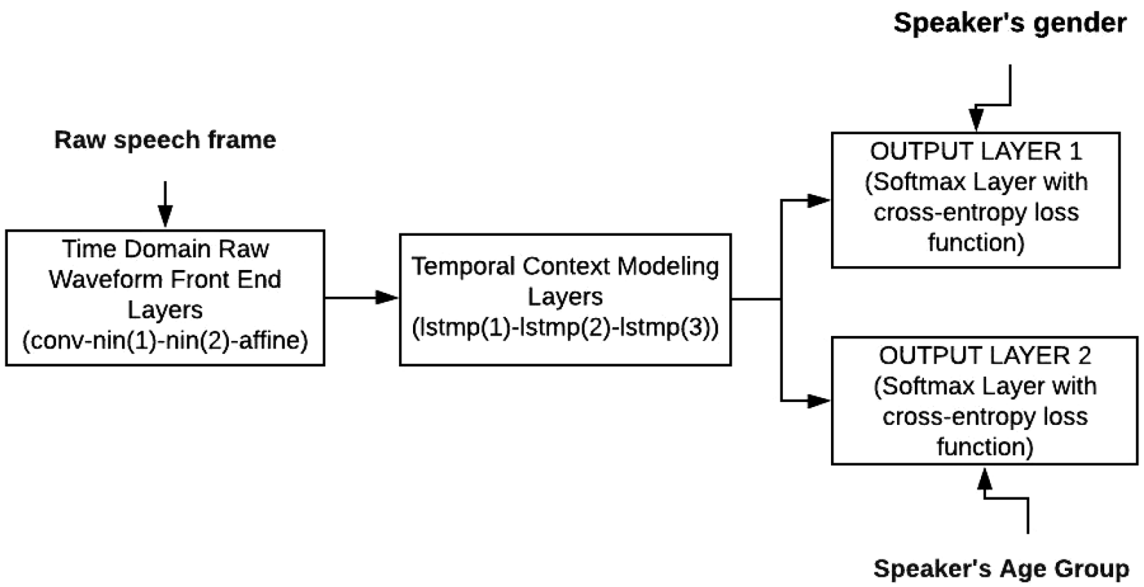


Fig. 1 Multi-task learning DNN set up for gender recognition. The raw waveform front end is shown in Fig. 2 and temporal context modeling layers are shown in Fig. 3

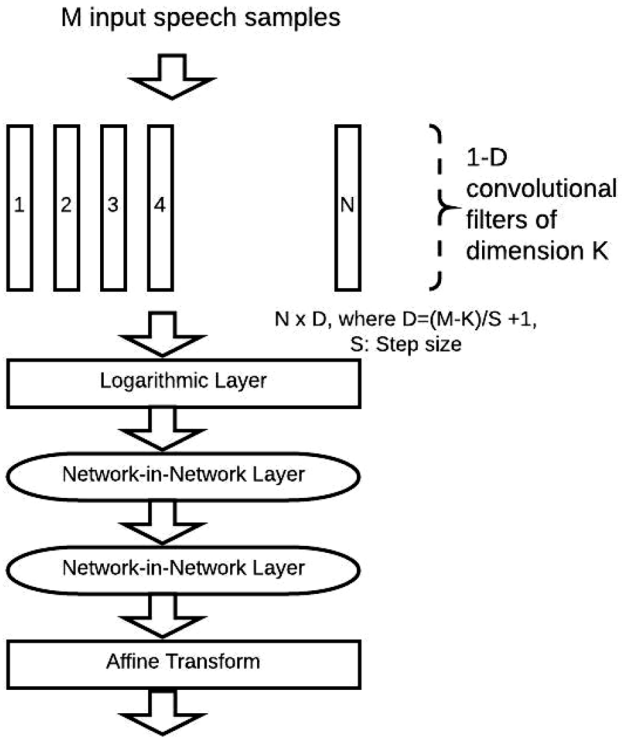
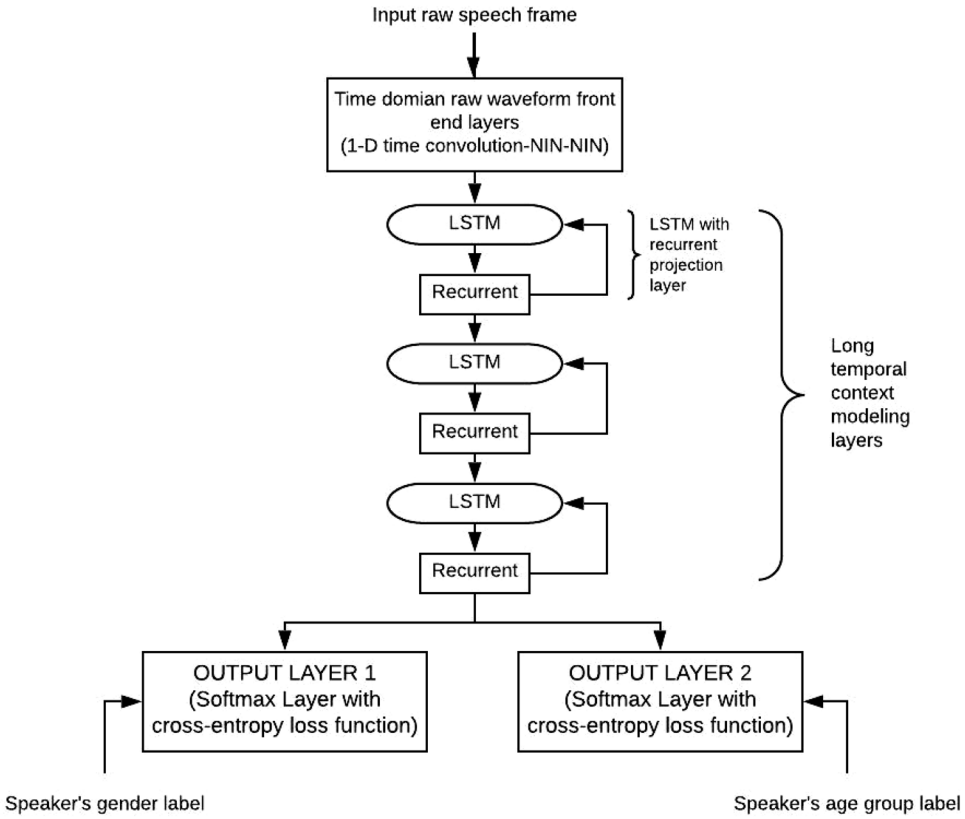


Fig. 2 Time domain raw waveform front end layers

2. Temporal context modeling layers as shown in the middle part of Fig. 3.
3. Output layers as shown in the bottom part of Fig. 3.

The front-end block is consisting of convolutional layer and two NIN layers and it takes raw speech frame as input. On the other hand, LSTMP layers are used for temporal context modeling which takes the feature representation discovered by the front end layers as input. However, all these layers are jointly trained and optimized with two separate output layers with gender and age targets respectively. Both the output layers are trained using softmax layer with cross entropy loss function. This gives the network freedom to model any distribution over output. Male and female gender is modeled as separate output classes in the primary output layer. Age of speaker is divided into seven groups: children, young female, young male, adult female, adult male, senior female and senior male. All these groups are modeled as seven classes in the auxiliary output layer. Individual layer components are described in detail in Sect. 5. Following here is a mathematical description of the proposed multi-task learning DNN in Sect. 4.1 and a step by step description of the complete algorithm is given in Sect. 4.2.

Fig. 3 Long temporal context modeling layers and output layers are shown which takes feature representation discovered by raw front end layers as input. The front end is shown in Fig. 2



4.1 Mathematical model of the proposed multi-task learning DNN

Mathematically, we can state the proposed problem of multi-task learning gender recognition as follows:

Let, $x \in X$ be a sequence of feature vectors (sequence of raw speech frames in the proposed work) extracted from a speech signal or utterance s through a window of length 40 millisecond (ms) and shift of 10 ms. The variable x can be expressed as,

$$x = (x_1, x_2, \cdots, x_T) \tag{1}$$

Here, T is the total number of frames in the utterance s or the length of the sequence and $x_i, i = 1, \cdots, T$ represents one speech frame of utterance s.

Let, $y \in Y$ is the corresponding gender labels, where y can be expressed as

$$y = (y_1, \cdots, y_T) \tag{2}$$

Here, $y_i \in 0, ..., K-1$, where $K = 2$ is the number of gender classes.

Let, $a \in A$ is the corresponding ageGender group labels, where, a can be written as

$$a = a_1, \cdots, a_T \tag{3}$$

Here, $a_i \in 0, ..., J-1$, where $J = 7$ is the number of ageGender group classes.

The multi-task learning objective is to find two functions $M : X \rightarrow Y$ and $N : X \rightarrow A$ that matches an input sequence x to the corresponding label sequence y and a respectively.

In this work, we propose to model these two functions $M : X \rightarrow Y$ and $N : X \rightarrow A$ through a stack of neural network layers consisted of one 1-dimensional convolutional layer ($conv$), two NIN layers ($nin^{(1)} - nin^{(2)}$), three LSTMP layers ($lstmp^{(1)} - lstmp^{(2)} - lstmp^{(3)}$) and two softmax layers ($soft^{(1)}$, $soft^{(2)}$). The softmax layers model the distribution over each output separately taking the output of $lstmp^{(3)}$ as input. The entire model is a chain of composite functions, each layer is a function which acts on the output of the previous layer. Every layer of the network transforms the input raw speech data into a new representation so that the mapping tasks becomes easier. In the proposed model, the first three layers ($conv - nin^{(1)} - nin^{(2)}$) generates a feature representation from the raw speech frame and these feature representation is used by the three LSTMP layers ($lstmp^{(1)} - lstmp^{(2)} - lstmp^{(3)}$) to model temporal dynamics of speech. These two independent block of layers are connected through an affine transformation layer ($affine$) and all layers are jointly trained. The output of all these layers are defined by the set of equations as in the following,

$$O_{conv} = f_{conv}(x) \tag{4}$$

$$O_{nin^{(1)}} = f_{nin^{(1)}}(|\log_{10}(O_{conv})|) \tag{5}$$

$$O_{nin^{(2)}} = f_{nin^{(2)}}(O_{nin^{(1)}}) \tag{6}$$

$$O_{affine} = f_{affine}(O_{nin^{(2)}}) \tag{7}$$

$$O_{lstmp^{(1)}} = f_{lstmp^{(1)}}(O_{affine}) \tag{8}$$

$$O_{lstmp^{(2)}} = f_{lstmp^{(2)}}(O_{lstmp^{(1)}}) \tag{9}$$

$$O_{lstmp^{(3)}} = f_{lstmp^{(3)}}(O_{lstmp^{(2)}}) \tag{10}$$

where $f(.)$ represents the overall function modeled by the layer mentioned in subscript and O represents the output of that layer, which is usually a matrix where number of row is equal to the number of speech frames and number of column is equal to the dimension of the layer. Primarily, the layer outputs are at frame level. The convolutional layer in the Eq. 4 takes sequence of raw speech frames x as input. We obtain the learned feature representation aggregated over the convolutional filters at the output of the second NIN layer as $O_{nin^{(2)}}$ in Eq. 6 which is next used by the LSTMP layers for long temporal context modeling. $O_{lstmp^{(3)}}$ is the output of the last LSTMP layer represented by $f_{lstmp^{(3)}}(.)$ due to an input which is the output of second LSTMP layer. Later, in Eqs. 19 through 26 we will show that, output of LSTMP layers are primarily an appended version of a LSTM cell and a recurrence with a feedback connection to the input of LSTM cell. For the simplicity of description, we show the equations computed by $conv$, nin and $lstmp$ in Sects. 5.1 and 5.2.

Thus, finally the mapping function $M : X \rightarrow Y$ and $N : X \rightarrow A$ can be expressed as,

$$M = f_{soft^1}(O_{lstmp^{(3)}}) \tag{11}$$

and

$$N = f_{soft^2}(O_{lstmp^{(3)}}) \tag{12}$$

This chain of composed functions is optimized to perform two different tasks. To fulfill such multi-task objective, the two functions are modeled through shared same set of layers ($conv - nin^{(1)} - nin^{(2)} - affine - lstmp^{(1)} - lstmp^{(2)} - lstmp^{(3)}$) i.e. primarily a mapping of multiple outputs from the same representation is performed. The model is trained with cross entropy loss objective applied at two softmax layers separately, which minimizes cross-entropy between the distributions represented by the reference labels and the predicted distribution. The cross entropy objective loss for output layer having gender label can be expressed as the negative log posteriors summing over all frames of training data as,

$$L_{gender} = -\sum_{t=1}^{N} \log u_t(y_t) \tag{13}$$

where $u_t(y_t)$ is the output of the softmax output layer for the gender class y_i, expressed as,

$$u_t(y_t) = P(y_t \mid x_t) = \frac{\exp\{h_t(y_t)\}}{\sum_K \exp\{h_t(y_t)\}} \tag{14}$$

Here, $h_t(y_t)$ is the activation of a linear affine layer corresponding to class y_i and N is the total number of frames in the training set.

Similarly, the cross entropy objective loss for output layer having ageGender group label can be expressed as,

$$L_{age} = -\sum_{t=1}^{N} \log v_t(a_t) \tag{15}$$

where $v_t(a_t)$ is the output of the second softmax layer for the ageGender group class a_i, expressed as,

$$v_t(a_t) = P(a_t \mid x_t) = \frac{\exp\{g_t(a_t)\}}{\sum_J \exp\{g_t(a_t)\}} \tag{16}$$

Here, $g_t(a_t)$ is the activation of a linear affine layer corresponding to class a_i.

Therefore, the multi-task objective loss sums the loss of each task as given

$$L_{mtl} = L_{gender} + L_{age} \tag{17}$$

All layers of the model are jointly trained through an iterative process and error is back-propagated for each layer through the network.

4.2 Steps of the proposed algorithm

A step by step description of the proposed system development algorithm is shown in Fig. 4 and it can be summarized as outlined below:

- Initially, we collect speech data (SRE 2008, OGI kids corpus and SRE 2010) with gender and age labels to train and test multi-task learning DNN. It involves a sub step where OGI kids corpus is divided into train and test partition. Details of all datasets are described in Sect. 6.1.
- Next, we combine SRE 2008 and train partition of OGI kids corpus to create DNN train data partition which contains speakers ranging from 4 to 84 years age. Each speech utterance have gender and age label associated with it. We create seven ageGender groups from the age labels and ageGender groups are used as label by the auxiliary output layer of the DNN.
- In the following step, we perform data augmentation by means of amplitude perturbation to increase the amount of data in the training set. Amplitude perturbation also mitigate the effect of low and high amplitude variations of speech over ages. Amplitude perturbation is applied by modulating the amplitude of the raw speech signal. Each recording in the training data has been scaled with a random variable drawn from an uniform distribution over [0.125, 2].
- We further combine SRE 2010 and test partition of OGI kids corpus to create a held out test data which contains speakers ranging from 4 to 82 years age.
- Next, we also collect test data for cross validation of DNN model across various dataset and environment (SRE 2000 and Switchboard cellular part 1).

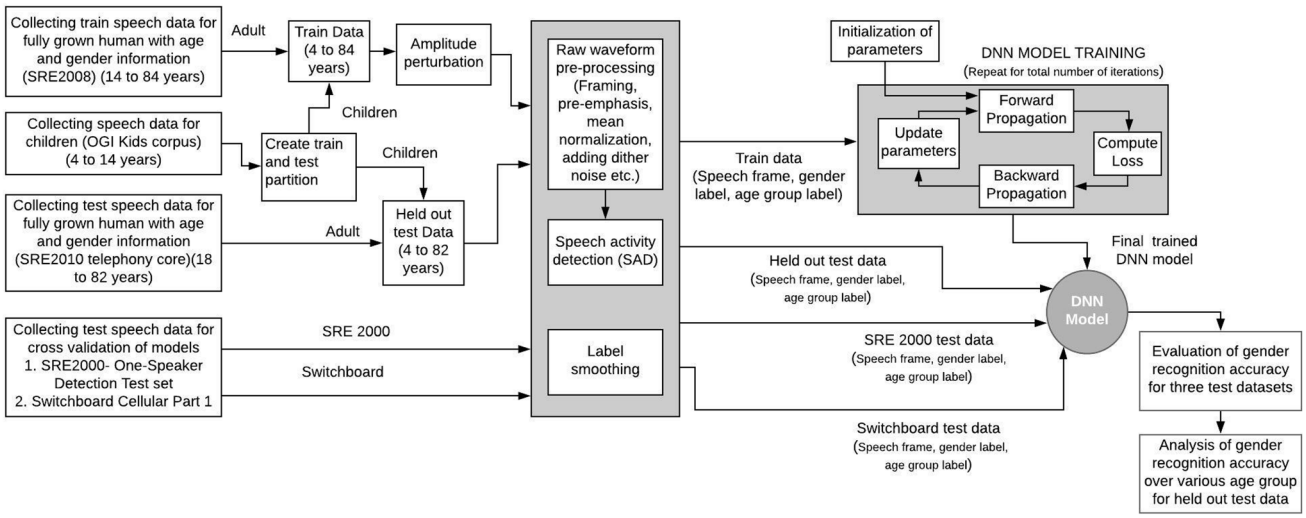


Fig. 4 Step by step description of the gender recognition system development algorithm. Layers of DNN model are shown in Fig. 3

- Raw speech waveforms are passed through a pre-processing block. It primarily creates raw speech frames from the wavefiles using frame size 40 ms and with a shift of 10 ms. We also removed DC offset by subtracting mean from waveform. Low energy dither is added to the windowed signal to avoid runs of digital zeros. All speech samples are considered at 8 KHz sampling frequency and quantized to 16 bit per sample. Speech samples which belongs to SRE 2000, SRE 2008, SRE 2010 and Switchboard datasets are already in 8 KHz and 16 bit sphere file format. These are converted to wav format during run time using sph2pipe tool provided by LDC through shell pipeline. OGI kids corpus samples which are in 16 KHz and 16 bit wav file format. We convert these sample using sox audio manipulation tool through shell pipeline during run time.
- Next, we perform speech activity detection (SAD) to create speech (1) and non speech (0) label for each of the speech frame based on energy thresholding on top of MFCC features. This 0/1 label corresponding to each raw speech frame is used to filter out non speech frames before creating a speech sequence (or speech chunk) to train the DNN.
- The algorithm avoids one hot encoding. Therefore, in the next step we perform label smoothing on the gender and ageGender group label used for DNN supervision, which helps the DNN to model any noise in the output distribution (mistakes in the data labels). Most datasets have some number of mistakes in the labels which can be harmful and therefore label smoothing is a widely used technique to explicitly model the noise on the labels. Label smoothing regularizes our model based on a softmax with k output classes by replacing the hard 0 and 1 classification targets (one hot encoding) with targets of $\frac{\epsilon}{k-1}$ and $1-\epsilon$, respectively, where ϵ is small constant ($\epsilon = 0.01$ for the DNN described in this work).
- The raw speech frame and corresponding labels are next used to train the DNN through Natural Gradient–Stochastic Gradient Descent (NG–SGD) algorithm (Povey et al. 2015) iteratively. Network training considerations are described in Sect. 6.4.
- The final model obtained from the training process is used for evaluation of the held out test data and other two datasets to obtain gender recognition accuracy.
- Finally, we perform analysis of the gender recognition accuracy in terms of ageGender group on the held out test data.

5 Feature learning and temporal context modeling in the proposed DNN

The neural network layers in the proposed multi-task DNN can be divided as feature learning layers, temporal context modeling layers and output layers. The layer configuration of feature extraction block is shown in Fig. 2 and temporal context modeling and output layers are shown in Fig. 3. Following here is a detail description of the components of each of these layers.

5.1 Feature learning layers

Extracting robust feature from speech signal which best represents information related to para-linguistic state of the speaker has been a critical problem. Most of the speech based systems use short term hand-crafted spectral and cepstral features based on fixed filters, such as MFCC or Mel filter-bank's output. However, using fixed filter may not be the most appropriate for minimization of classification error in a multi-task learning DNN setup. In such architectures, the learning and generalization of two tasks depends on the shared hidden representations. Evidences of relevant features for the two tasks are learned throughout the training process of the network. Therefore, in this work we exploit the feature learning power of DNN for multi-task learning purpose. We use front-end block as shown in Fig. 2 which attempts to learn filters within the DNN (Ghahremani et al. 2016; Sarma et al. 2018).

Such front-end block has a 1-dimensional time convolution layer, which operates on 50 ms raw signal with step size 1.25 ms. Five consecutive raw speech frames, x_{-2}, x_{-1}, x_0, x_1 and x_2 are concatenated and passed to the input layer of DNN. The 1-dimensional temporal convolutional layer has N number of K-dimensional filters. It takes M samples of the raw speech waveform and convolves with the N filters and generates a $C = N \times D$-dimensional output, where

$$D = \frac{(M-K)}{S} + 1 \tag{18}$$

Here, S is the step size taken along input axis. Thus, output of convolution is sub-sampled by step size rate which significantly reduces the computation time. In this work, 100 filters of 31.25 ms dimension is applied to 50ms speech with a filter shift (step size) of 1.25 ms. Next, absolute logarithm of filter outputs are computed using a logarithmic layer. Filter outputs are aggregated using two trainable Network-in-Network (NIN) non linearity layers introduced by Ghahremani et al. (2016). The NIN is a many-to-many non-linearity, where the input to NIN is initially passed through a Rectified Linear Unit (ReLU) layer followed by a micro-neural network block. The micro neural network block has an affine transformation block of size $m \times h$ ($m < h$) and the output is subsequently passed through another layer with ReLU activation. The output of ReLU layer is passed through a second affine transformation block of size $h \times n$ ($h > n$) followed by another ReLU layer. Thus, the entire NIN layer has two micro neural network blocks with shared ReLU layers.

Affine layer simply implements a linear function. Such NIN layers are capable of learning local connectivity within non overlapping region of speech. Such advantage can not be achieved in the case of max pooling based dimension reduction traditionally applied after convolution filters (Ghahremani et al. 2016).

5.2 Temporal context modeling layers

Speech is a time-varying signal with complex correlations between acoustic events within certain range of times. Therefore, while implementing DNN based learning to speech signal, the major challenge is to model the temporal dynamics. Particularly for para-linguistic information extraction, it is important to preserve long temporal context since such information perhaps mostly lies in a longer span of time. There could be two approaches: first is to capture the long term temporal behaviour during feature extraction [eg. TRAP feature (Hermansky and Sharma 1998)] and present it to the DNN. The second approach is to make DNN able to model long term dependencies from short term frames. The general structure of feed-forward neural network is not capable of learning long term dependencies due to its drawback of having separate parameters for each input unit (absence of parameters sharing across various parts of the model). However, over time architectures like recurrent network and TDNN evolved, which effectively models time varying data and also led to some major success in the area of speech recognition and other speech based task. Therefore, in this work we decide to compare TDNN with LSTM network for temporal context modeling and consider the one which provides best accuracy for the present problem of multi-task learning DNN.

TDNN captures long range temporality from short-term feature representations using a hierarchy of layers and a time delay arrangement (Waibel et al. 1989). The basic unit of the neural network is modified by introducing delays TD_1 through TD_n in the TDNN. For example, T inputs of one unit would be multiplied by the weights of surrounding n units plus the present unit. Thus the TDNN units has an ability to relate present instant of time with the history of events specified by the delay. The modern architecture of TDNN (Peddinti et al. 2015) used in this work has a sub-sampling architecture, where each layer is associated with a hyper-parameter called layer context. Layer context primarily defines, range of input units required to compute an output activation, at one time step. For example, at a particular layer, if input frames $t-2$ through $t+2$ are spliced together to compute output for current time instant, then context for the layer is written as $\{-2,-1,0,1,2\}$. We perform preliminary experiments with TDNN layer based DNN using mel filter bank output as feature as described in Sect. 6.3. It has been observed that TDNN performance is little below than LSTM based recurrent layers.

Therefore, in the proposed multi-task learning DNN, we have used recurrent modeling of temporal context using LSTMP layers (Fig. 5.2) for long temporal context modeling to optimize gender identification objective. Unlike traditional feed forward DNN, the recurrent layers have feedback connections between units. This creates an internal state of the network, which acts like a memory unit. For a few instants of time it persists with information circulation and extraction of relevant information and uses that contextual knowledge while processing the current instant. Practically, RNN processes chunk of input frames starting with the current frame and use the delay specified during training to generate contextual processing. The biological motivation behind RNN is that, human brain never starts to think everything from scratch in every second (Hochreiter and Schmidhuber 1997). Understanding of every word is based on the previous word's knowledge. This makes RNN a very special kind of DNN, which exhibits dynamic temporal behaviour. Thus, recurrent learning is a better approach to model temporal sequences like speech than other available solutions. The LSTM cell based recurrent projection layers (LSTMP) (Sak et al. 2014), used in this work have memory cells and cell state carries all information along the network, with very minor changes. The flow of information is controlled by some gates, which are basically some multiplicative units. To maintain the flow of description, we briefly reviews the equations computed inside LSTMP memory cell below.

A LSTMP layer computes a mapping from an input sequence $h = (h_1, ..., h_T)$ to an output sequence $\hat{h} = (\hat{h}_1, ..., \hat{h}_T)$ by calculating the network unit activation using the following equations iteratively from $t = 1$ to T:

$$i_t = \sigma(W_{ix}h_t + W_{im}r_{t-1} + W_{ic}c_{t-1} + b_i) \tag{19}$$

$$f_t = \sigma(W_{fx}h_t + W_{fm}r_{t-1} + W_{fc}c_{t-1} + b_f) \tag{20}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{cx}h_t + W_{cm}r_{t-1} + b_c) \tag{21}$$

$$o_t = \sigma(W_{ox}h_t + W_{om}r_{t-1} + W_{oc}c_t + b_o) \tag{22}$$

$$m_t = o_t \odot tanh(c_t) \tag{23}$$

$$p_t = W_{pm}m_t \tag{24}$$

$$r_t = W_{rm}m_t \tag{25}$$

$$\hat{h}_t = (p_t, r_t) \tag{26}$$

where W terms denote weight matrices (e.g. W_{ix} is the matrix of weights from the input to the input gate), b terms denote bias vectors (b_i is the input gate bias vector), σ is the logistic sigmoid function and i, f, o and c are respectively the input gate, forget gate, output gate and cell activation vectors. The $\odot$ is the element-wise product of the vectors. The linear projection layer projects the memory cells outputs m_t (Eq. 25) to a lower-dimensional vector r_t by a linear transform W_{rm} and r_t will be fed as inputs for the next time step. The dimensions of projection and the recurrence in LSTM layers are one quarter of the cell dimension. We found 512 cell dimension optimal for the current gender identification task with the recurrence of dimension 128 and the LSTM output of dimension 256. The LSTMs operates with a recurrence that spans 3 time steps. Three consecutive LSTM with recurrent projection layers are used in the present DNN set up. However, these hyper-parameters are selected based on a few initial baseline experiments as described in Sect. 6.3.

6 Experimental details

This section describes the experimental details and result. All our experiments are done using Kaldi toolkit (Povey et al. 2011). Following here is a detail description of experimental considerations.

6.1 Database and performance matrices

The objective of this work is to validate the usefulness of age information to classify gender from speaker's belonging to different age group. Therefore, we combine fully grown human's and children's speech from different datasets to create our train and test partition.

Our train set uses fully grown human's speech from *NIST SRE 2008 dataset* (2008). We use wave files from 10 *s* and *short*2 train conditions and 10 *s* and *short*3 test conditions of *SRE 2008*. Thus we obtain a total of 1141 adult speakers with age ranging from 14 to 84, where 419 speakers are male and 722 speakers are female. Further, to obtain children's speech sample, we create a train and test partition of 1093 speakers of *OGI Kids' Speech corpus* (Shobaki et al. 2000). The entire dataset has children from Kindergarten through grade 10 which is roughly from age 4 to 14 years. i.e total 11 different ages. For our test data we pick speech utterances from 4 male and 4 female speakers from each age. Thus the test partition has 88 speakers, whereas rest 1005 speakers are used for training. We combine this *OGI* train partition (1005 speakers) with 1141 adult speakers of *SRE 2008* to create the final train set for our gender identification system. Thus, the train set has 2146 speakers of 4 to 84 age and a total speech of 591.23 h (349.06 h female data and 242.16 h male data). We further perform amplitude perturbation on this training set to double the training data amount.

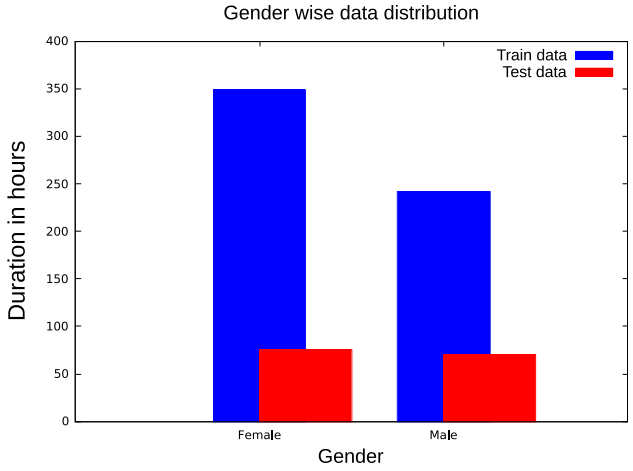
On the other hand, *SRE 2010 telephony core dataset* (2010) and the test subset of *OGI Kids* (mentioned above) are considered as our held out test data, totaling 146.21 h of speech. *SRE 2010 telephony core dataset* have 410 speakers with age information which ranges from 18 to 82 year. Therefore, this makes an appropriate test data for the present study. Further the conversational telephone speech of *SRE 2010* is recorded including some variety like phone calls made with high, low, and normal vocal effort, one channel using room microphone channel etc. This makes *SRE 2010* dataset more real life oriented. The train and test data distribution in terms of total duration of speech is shown in Fig. 5 age wise and gender wise.

Apart from this, we also produce our gender recognition evaluation results for *SRE Evaluation* (2000) and *Switchboard Cellular Part 1* datasets (https://catalog.ldc.upenn.edu/LDC2001S13) to validate the suitability of the models across different datasets and environment. *One-Speaker Detection Test set* of *SRE 2000 dataset* is used which has 1061 speakers (596 male speakers, 563 female speakers) and total 9.08 h telephone speech data. *Switchboard Cellular Part 1* is a GSM cellular phone audio dataset, consists of approximately 109 h of English telephone conversations from 254 speakers (129 male speakers, 125 female speakers) under varied environmental conditions.

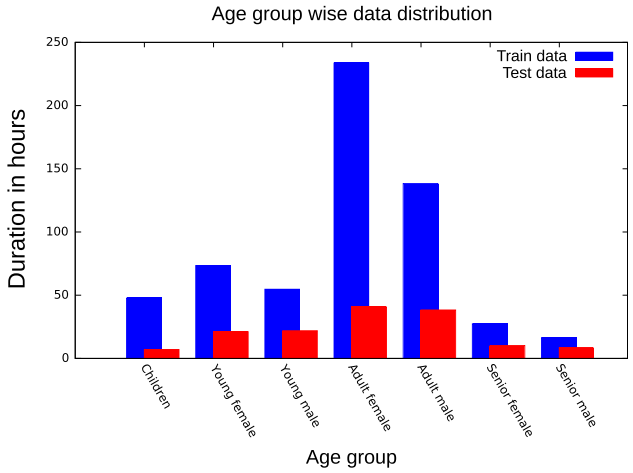
The performance of the gender identification tasks are reported using two parameters, weighted accuracy (WA) which is the overall classification accuracy and unweighted accuracy (UA) which is the average recall over the gender categories. Apart from that individual percentage of correct classification for male (%M) and female (%F) are also reported.

6.2 MFCC feature based GMM baseline

We design MFCC feature based Gaussian mixture model (GMM) baseline to make a comparative analyses of proposed DNNs. We consider 20 ms frame window and 10 ms shift to compute 19 MFCC augmented with energy, 20 delta and 20 acceleration coefficients, creating 60-dimensional MFCC vectors and these are used to train GMMs. Speech segments are passed through energy SAD module to compute 0/1 sequence which is used to filter out the non speech frames. The GMM module is basically a classification algorithm based on GMMs log-likelihood ratio. We train two GMM models with frame level MFCC features for male and female. Gender classification for a test utterance X is done by comparing the log-likelihood of the two models $L(X)_{male}$ and $L(X)_{female}$ respectively. We compute the log-likelihood ratio for classification and a confidence measure using the following two equations,



(a) Gender wise data distribution in train and test data



(b) Age wise data distribution in train and test data

Fig. 5 Distribution of data in train and test set

$$C(X) = \mid 2P(X) - 1 \mid \tag{27}$$

where $P(X)$ is given by,

$$P(X) = \sigma(L(X)_{male} - L(X)_{female}) \tag{28}$$

where $\sigma(.)$ is the sigmoid function.

6.3 Single-task learning DNN baseline with mel filter bank outputs

A baseline DNN has been designed where mel-filter bank's outputs (40-MFCC hereafter) are used as feature. The process of generating these features are similar to that of MFCC except that the last cepstral truncation step is not used. Essentially log filter bank energies for all mel filters are kept. Thus, the feature vector is a 40 dimensional vector corresponding to 40 filters. These features are equivalent to mel-filter bank's output (Povey et al. 2015; Peddinti et al. 2015). But since MFCC is less correlated, these feature are more easily compressible. Features are extracted with a frame-length of 25 ms and 10 ms overlap. Cepstral mean subtraction is performed over a window of 6 s.

Here, two different long temporal context modeling approaches were experimented. First, using temporal convolution in the form of TDNN layers and secondly, using only LSTMP layers. We have provided a detail description of how LSTMP layer and TDNN layer functions in Sect. 5.2. The hyper-parameters which defines the structure of TDNN are the input contexts of each layer required to compute an output activation, at one time step. Context of TDNN layers are specified in terms of the splicing indices (Peddinti et al. 2015). It defines the temporal convolution kernel input at each layer. TDNN models long term temporal dependencies through a feed-forward DNN approach unlike LSTMP based recurrent layers where feedback connections are used. On the other hand, hyper-parameters which defines LSTMP layers are the dimension of LSTM cell, dimensions of recurrence and non recurrence projection and the delay time. For this baseline experiment 40 dimensional feature vector is passed to the TDNN or LSTMP layers as input as shown in Fig. 6. We investigate various width and height of the DNN, changing number of layers and dimension of the layer. A softmax layer with cross-entropy loss is used at the end of TDNN or LSTMP layers, where genders are used as target.

The results of experiments which uses TDNN layers only are shown in the first part of Table 2. The TDNN configuration has five TDNN layers stacked, with layer context as shown in Table 1. Various output dimension of the layers have been investigated. The best WA obtained from the TDNN based single-task learning DNN is 75.09%, using 5 layers of TDNN with output dimension of 512. On the other hand, number of LSTM layers are varied and lower parts of the Table 2 shows the results. Three LSTM with cell dimension 512, recurrence dimension 128 and output dimension 32 gives the best pair of WA and UA i.e. 77.30% and 77.35%. The choice of network configuration should be primarily a trade-off between training time which depends on the trainable parameters and accuracy. We observe that, in case of LSTMP networks we achieve accuracy of 77.30% using 3 layers with lower cell dimension like 512 and 2649858 number of total trainable parameters. However, in case of TDNN networks we achieve 75.09% accuracy using 5 TDNN layers of 512 dimension and the total trainable parameters of the network is 3456514. This shows that, even with more number of trainable parameters, we obtain low accuracy in TDNN compared to LSTMP. TDNN networks need more layers compared to the LSTMP to fit this particular problem and obtain an optimized robust model. Therefore, from the perspective of computational cost, network training time and accuracy LSTMP network is a better

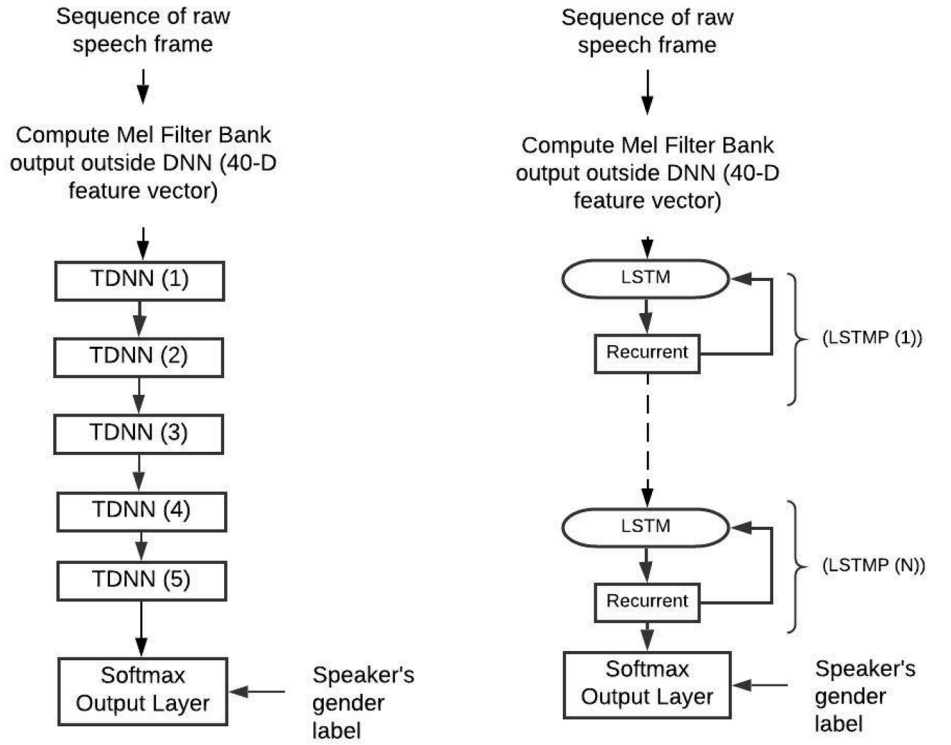


Fig. 6 Architecture of single-task learning DNN where 40-MFCC is used as input feature

(a) TDNN based single-task learning DNN, which uses layer context as shown in Table 1 and layer output dimension varies as shown in Table 2.

(b) LSTMP based single-task learning DNN. Total number of LSTMP layers and dimensions varies as described in the results of Table 2

Table 1 Layer wise context of TDNN based single-task DNN model for gender identification

Layer	Context	Layer-type
1	[−1, 0, 1]	TDNN
2	[−2, 0, 2]	TDNN
3	[−3, 0, 3]	TDNN
4	[−6, 0, 6]	TDNN
5	[−12, 0, 12]	TDNN

choice than TDNN. Further, in case of LSTMP networks, we observed that with increasing number of parameters in the network, the accuracy reduces. This provides a hint that training bigger network in case of TDNN also may not be much helpful. Therefore, with the background of these initial experiments, raw waveform based DNN experiments are performed, where the same LSTMP configuration for temporal context modeling is used.

6.4 Multi-task learning DNN training considerations

Networks are trained using standard NG-SGD algorithm of Kaldi systems (Povey et al. 2015, 2011), which uses a

Table 2 Results of gender identification using 40-MFCC in single-task DNN on held out test data (SRE 2010 + OGI Kids test part)

Model	Number of layers for temporal context modeling	Dimension	Total number of parameters	WA	UA
TDNN	5	128	274,306	74.23	74.43
	5	512	3456,514	75.09	75.34
LSTM	3	Cell = 128, recurrence = 32, output = 64	244,674	60.42	60.67
	5	Cell = 128, recurrence = 32, output = 64	361,282	69.02	69.05
	3	Cell = 512, recurrence = 128, output = 256	2649,858	77.30	77.35
	5	Cell = 512, recurrence = 128, output = 256	4492,546	70.25	70.22
	3	Cell = 1024, recurrence = 256, output = 512	9756,162	68.54	68.46

sophisticated parameter averaging techniques to facilitate neural network training parallely distributed across several GPUs or CPUs. The parameter-averaging method of NG-SGD, first allows multiple SGD processes running on separate GPUs with different randomized subsets of the training data. After each GPU processes a fixed number of training examples, parameters across all the jobs are averaged and redistributed the result to the jobs. This is repeated until all the data for a specified number of epochs are processed. Thus, the number of epochs is determined in advance (stopping criterion is not used), which is considered as a hyper-parameter in the experiments reported here. Selection of number of epochs depends to some extent on the size of mini-batch i.e. total number of chunks used per mini-batch. We observe validation data accuracy based on number of epochs and number of chunks per mini-batch and thus arrive at an optimal pair. The learning rate of the training is set to be gradually decrease from 6×10^{-4} to 6×10^{-5} over the course of 30 epochs and a fixed mini-batch size (100 chunks per mini-batch) is used.

We use per frame dropout using the dropout schedule method described in Cheng et al. (2017) where entire vector is forced to be zero or one. The dropout schedule is expressed as a piece wise linear function on the interval [0, 1], where $f(0)$ gives the dropout proportion at the start of training and $f(1)$ gives the dropout proportion after seeing all the data. A dropout schedule of the form 0, 0@0.20, p@0.5, 0@0.75, 0 is used in this setup, where p is 0.3 in the results reported here. Thus, the dropout probability is 0 at $f(0)$, 0 at $f(0.2)$, 0.3 at $f(0.5)$, 0 at $f(0.75)$ and 0 at $f(1)$.

We train DNNs computing objective per frame i.e label is repeated for every frame of the speech chunk. Thus, outputs of DNN are at frame level and hence to obtain time aggregation over the frames of a chunk, we do some post processing outside the network. Basically, we average frame posteriors to get a segment level aggregate from the frame level posteriors. At the end, we give some extra left context at the time of decoding (while running the DNNs to evaluate the test examples), which provides flexibility to the network regarding number of frames it sees in addition to what was provided during training the model and we evaluate the model several times to tune this length of decode time context.

6.5 Results and analysis of DNNs trained using raw waveform

At the beginning of this investigation, we evaluated the GMM based gender identification system for three datasets, *SRE 2000 one Speaker detection task* test segments, *Switchboard Cellular Part 1* segments and our held out test data (combined *SRE 2010 core* data and test partition of *OGI Kids* corpus). The results of all three tests is shown in the first part of Table 3. It can be observed that for *SRE 2000* and *Switchboard*, we obtain above 96% WA, whereas the performance degrades for the held out test data (86.7%), where speakers from all age groups (children to seniors) are present..

After that we performed the same experiment on our raw waveform based single-task learning DNN set up, which is trained to classify two genders of speakers using a softmax output layer with cross entropy objective. The configuration of this DNN is similar to that in Fig. 1, except that it has only one output layer with gender targets. However, similar degradation of WA and UA is observed on the held out test data, as shown in second part of the Table 3.

We assumed that the reason for low accuracy on our held out test dataset is perhaps the wide range of speaker's age on this dataset. This leads us to the design of multi-task learning

Table 3 Results of gender identification using GMM, Single-task learning DNN and multi-task learning DNN model on all test data

Model	Dataset	WA	UA	F%	M%
GMM	SRE 2000	98.68	98.68	98.81	98.56
	Switchboard Cellular Part 1	97.7	97.69	97.72	97.67
	SRE 2010	77.98	78.11	71.87	84.34
	OGI Kids test part	62.41	61.66	99.46	23.86
	Held out test data (SRE 2010 + OGI Kids test part)	86.739	86.8	85.17	88.42
Single-task DNN	SRE 2000	98.68	98.7	99.2	98.2
	Switchboard Cellular Part 1	97.04	96.91	98.13	95.69
	SRE 2010	88.05	88.18	84.76	91.6
	OGI Kids test part	77.98	78.11	71.87	84.34
	Held out test data SRE 2010 + OGI Kids test part	87.89	87.77	84.73	89.49
Multi-task DNN	SRE 2000	98.39	98.41	98.81	98.02
	Switchboard Cellular Part 1	96.89	96.68	98.61	94.74
	SRE 2010	91.13	91.09	92.2	89.97
	OGI Kids test part	76.35	76.44	71.87	81.0
	Held out test data (SRE 2010 + OGI Kids test part)	90.7	90.65	91.81	89.49

DNN where we provide speaker's age as an additional target along with gender to the learning process of DNN. We divide the age of speaker to seven distinct groups as mentioned previously (children, young female, young male, adult female, adult male, senior female and senior male). Thus the network is allowed to discriminate gender related feature according to the age of the speaker. Significant improvement on the accuracy using multi-task learning set up has been observed. The performance of the multi-task learning DNN is shown in Table 4 for the entire datasets and files which are longer than 10 s duration. It can be observed that for both set of speech files DNN based multi-task learning improves gender recognition performance. For files longer than 10 s, 3.97% and for all files 11.59% absolute improvement is observed from multi-task learning over GMM.

The training objective curves of single-task and multi-task learning DNN are shown in Fig. 7. Here, multi-task learning DNN model is trained for 10 epochs with 32 number of chunks per mini-batch and single-task training DNN model is trained for 30 epochs with 32 number of chunks per mini-batch. It can be observed from the plots that multi-task learning model converges faster and final objective value is better than the single-task learning model.

Table 4 Results of gender identification, multi-task learning DNN on the held out test data

Duration of speech	Model	WA	UA	F%	M%
≥ 10s	GMM	86.739	86.8	85.17	88.42
	DNN STL	87.89	87.77	84.73	89.49
	DNN MTL	90.7	90.65	91.81	89.49
All data (1.21 s+)	GMM	69.01	68.35	95.53	41.16
	DNN STL	80.5	80.71	75.24	86.19
	DNN MTL	80.6	80.64	78.67	82.62

The last part of Table 3 shows the performance of the multi-task learning model on all other test datasets. It can be observed that for all other datasets multi-task learning DNN performs equally well with GMM and single-task learning DNN, at the same time improves at the held out test dataset which contains speaker's from age.

We further investigated the performance of the multi-task learning DNN based on the age groups, which is shown in Tables 5 and 6 and compared with the single-task and GMM counter parts. As can be observed from the last part of Table 5, multi-task learning DNN model is more generalized across age of speaker and hence improvement in terms particular age group like young, adult and senior is visible. On the other hand, the model is more generalized across gender as well, which can be observed from individual recognition rate of male and female. Further, significant improvement has been observed on the WA and UA of children data using DNN model (61.9 → 82.14 WA and 55.55 → 81.59 UA). Both single-task and multi-task model gives equal performance for 10sec above children wave files.

In Table 7, we have compared all classifiers designed as part of this work in terms of relative reduction of error rate (RRE) computed using the following equation,

$$RRE = \frac{er_b - er_p}{er_b} \times 100 \tag{29}$$

where er_b is the error rate in the baseline classifier (reference) and er_p is the error rate in the new classifier. Both er_b

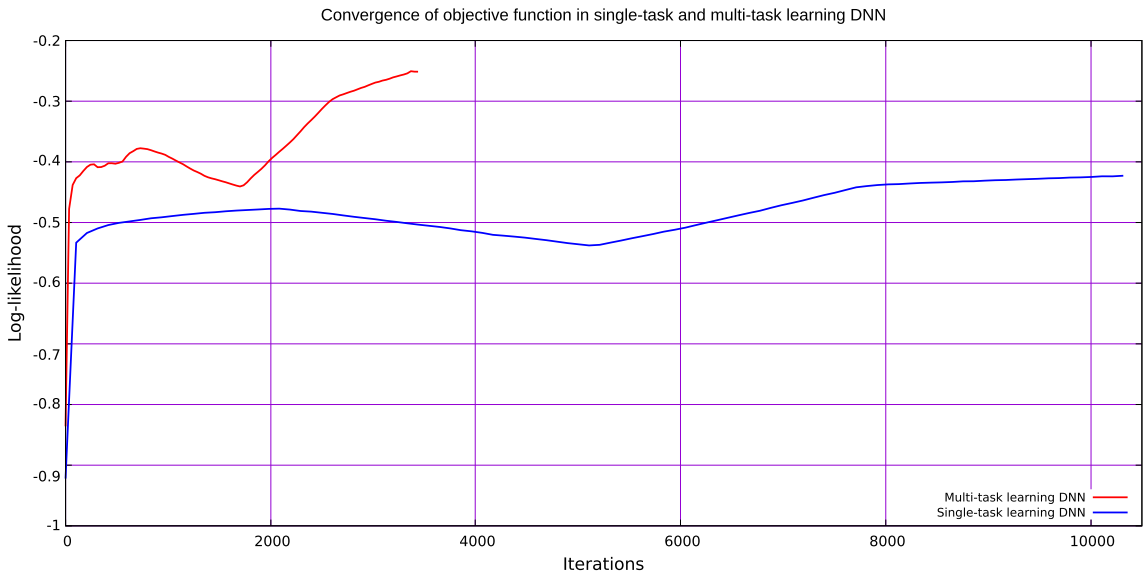


Fig. 7 DNN training objective curve for gender recognition

Table 5 Distribution of accuracy over age group (speech duration ≥ 10s) on the held out test data

Model	Age group (age range)	WA	UA	F%	M%
GMM	Children (4–14)	61.9	55.55	100	11.11
	Fully grown (15–80)	87.99	88.12	84.34	91.89
	Young (15–24)	89.15	89.09	84.8	93.38
	Adults (25–54)	87.6	87.72	84.29	91.15
	Seniors (55–80)	86.97	87.5	83.6	91.39
Single-task DNN	Children (4-14)	82.14	81.59	85.41	77.77
	Fully grown (15–80)	87.99	88.17	84.69	91.64
	Young (15–24)	88.95	88.89	84.8	92.99
	Adults (25–54)	87.71	87.81	84.71	90.92
	Seniors (55–80)	87.44	87.91	84.42	91.39
Multi-task DNN	Children (4-14)	82.14	81.59	85.41	77.77
	Fully grown (15–80)	91.13	91.09	92.17	90.02
	Young (15–24)	92.89	92.9	93.2	92.6
	Adults (25–54)	90.91	90.85	88.93	92.76
	Seniors (55–80)	87.9	87.93	88.17	87.7

Table 6 Distribution of accuracy over age group (speech duration 1.21 s+) on the held out test data

Model	Age group (age range)	WA	UA	F%	M%
GMM	Children (4–14)	58.57	57.50	99.40	15.60
	Fully grown (15–80)	89.15	89.17	88.04	90.30
	Young (15–24)	90.98	91.0	92.62	89.39
	Adults (25–54)	87.60	87.72	84.29	91.15
	Seniors (55–80)	86.97	87.50	83.60	91.39
Single-task DNN	Children (4–14)	75.95	76.11	69.87	82.35
	Fully grown (15–80)	89.51	89.6	85.63	93.56
	Young (15–24)	91.56	91.5	86.79	96.21
	Adults (25–54)	87.71	87.81	84.71	90.92
	Seniors (55–80)	87.44	87.91	84.42	91.39
Multi-task DNN	Children (4–14)	74.09	74.2	69.78	78.62
	Fully grown (15–80)	91.88	91.89	91.52	92.264
	Young (15–24)	93.57	93.54	91.26	95.833
	Adults (25–54)	90.91	90.85	92.76	88.938
	Seniors (55–80)	87.9	87.93	87.704	88.172

and er_p are computed using the WA of the respective classifier mentioned in previous sections and subtracting it from 100. In Table 7, we consider one classifier as baseline reference at a time and another as a new classifier and compute the relative error reduction rate for six important cases as mentioned below:

1. TDNN based single-task learning DNN versus LSTM based single-task learning DNN trained using 40-MFCC as feature with RRE 8.87%.
2. LSTM based single-task learning DNN trained using 40-MFCC as feature versus LSTM based single-task learning DNN trained using raw speech with RRE 46.65%.
3. GMM trained using MFCC feature versus LSTM based single-task learning DNN trained using raw speech with 8.67%.
4. GMM trained using MFCC feature versus LSTM based multi-task learning DNN trained using raw speech with 29.91%.
5. LSTM based single-task learning DNN trained using 40-MFCC as feature versus LSTM based multi-task learning DNN trained using raw speech with RRE 59.03%.
6. LSTM based single-task learning versus multi-task learning DNN, both trained using raw speech with 23.20%.

It is clearly visible from the results of last three cases of the Table 7, that error rate reduces significantly with the Multi-task learning DNN architecture, compared to GMM, hand-crafted feature based Single-task learning DNN and learned feature based single-task learning DNN .

7 Summary of findings

In this work, we have reported how speaker's age can be used as an additional information to improve gender recognition accuracy. Design of a multi-task learning based DNN architecture has been described to prove our hypothesis. The proposed multi-task learning DNN uses age of speaker as target in an auxiliary output layer and optimize gender recognition objective minimizing the total loss of both the tasks.

In order to exploit the feature learning and discriminating ability of multi-task learning process of DNN, we used raw speech waveform in this work. A 1-D time convolutional layer and NIN layer based front-end block has been used to learn filters within the network from time domain waveform as described in Sect. 5.1. These layers are jointly trained with the speech temporal modeling parts of the DNN to optimize gender classification objectives. We use LSTMP based recurrent learning to preserve long temporal context of speech as described in Sect. 5.2. However, we have performed some baseline experiments using 40-MFCCs as feature to single-task learning DNN as described in Sect. 6.3. The selection of LSTMP is done based on these initial experiments where we compare LSTMP with TDNN layers and observed that LSTMP provides better accuracy (75.09% → 77.30%) with

Table 7 Relative reduction of error rate (RRE) for various classifier considering one as baseline at a time

Sl. no	Baseline classifier			New Classifier			RRE (%)
	Feature/input	Model	WA (%)	Feature/input	Model	WA (%)	
1	Mel-filter bank output	Single-task DNN with TDNN layers	75.09	Mel-filter bank output	Single-task DNN with LSTM layers	77.30	8.87
2	Mel-filter bank output	Single-task DNN with LSTM layers	77.30	Raw speech	Single-task DNN with Raw front end+LSTM layers	87.89	46.65
3	MFCC	GMM	86.73	Raw speech	Single-task DNN with Raw front end+LSTM layers	87.89	8.67
4	MFCC	GMM	86.73	Raw speech	Multi-task DNN with Raw front end+LSTM layers	90.7	29.91
5	Mel-filter bank output	Single-task DNN with LSTM layers	77.30	Raw speech	Multi-task DNN with Raw front end+LSTM layers	90.7	59.03
6	Raw speech	Single-task DNN with Raw front end+LSTM layers	87.89	Raw speech	Multi-task DNN with Raw front end+LSTM layers	90.7	23.20

a relative error rate reduction of 8.87%. On the other hand, in the subsequent experiments we observe relative error rate reduction of 46.65% using front-end layers in a single-task learning LSTMP-DNN compared to that of using 40-MFCC as feature (WA 77.30% → 87.89%). This leads us to design multi-task learning DNN using front-end layers for feature extraction and LSTMP layers for long temporal context modeling.

The multi-task learning DNN have two consecutive output layers. The primary output layer has two gender classes as target and the auxiliary output layer have seven ageGender groups as target. We observe that the proposed multi-task learning DNN set up generalizes better across age and gender of speaker compared to single-task DNN and GMM counter parts. We achieve significant improvement on gender recognition accuracy from the proposed DNN over GMM model (11.59% and 12.29% absolute improvement on WA and UA respectively). We have also evaluated the performance of all three set ups (GMM, single-task learning DNN and multi-task learning DNN) in case of other standard datasets like SRE 2000, Switchboard etc. to validate the universality of the model across different test sets. From the experimental results it can be clearly observed that gender recognition accuracy for *SRE 2000* test data is 98.39%, for *Switchboard Cellular Part 1* it is 96.89%, for *SRE 2010* test data it is 91.13%, for *OGI kids corpus* accuracy is 76.35% and for the held out test data accuracy is 90.7%. So it is clear that the accuracy is pretty high compared to many previous works on gender recognition like Li et al. (2010, 2013) Kumar et al. (2016) Kabil et al. (2018) etc. However, this is not an absolute comparison, since the test datasets used by various works are different. In this work, we use four standard datasets provided by prestigious Linguistic Data Consortium (LDC) as test data and studied how gender recognition performance degrades when tested with datasets containing children and senior persons' speech instead of only adults' speech. We provide a feasible solution to such situation by designing a classifier which acquires adequate learning regarding the speakers' age while attempting to discriminate gender. The proposed multi-task learning DNN classifier learns features which are supported by both gender and age objective, thus performs better on gender task. We are not aware of about any such previously reported work where interrelation of para-linguistic task has been investigated, such as the age dependency of gender discrimination process. We propose a system which takes advantage of such dependency of multiple para-linguistic tasks and improves at one task. Further, to the best of our knowledge, no previous works have reported the use of multi-task learning DNN set up for gender recognition using raw speech waveform as input and learning feature representation within the network. The present work can be extended to more DNN architectures like Time Delay Neural Network (TDNN) and attention mechanism using the methodology adopted here. We are also performing experimental work on age group classification using similar multi-task learning DNN setup. Although we have not yet observed improvement on age using gender information, we aim to do more investigation on this direction in our future work.

8 Conclusion

The objective of this study is to investigate if age of speaker can be additional information to improve gender recognition performance. From common sense it can be assumed that humans psychology to discriminate gender from speech information somewhere depends on age of the speaker. Therefore, we attempted to make the specially configured DNN model be aware about age of the speaker, while learning to discriminate speaker's gender. Experiments performed as part of this work have proven this hypothesis. A novel

multi-task learning DNN has been designed which consideres age and gender as two separate targets in two different output layers and minimizes the total objective loss. Learning with age auxiliary targets, the gender recognition accuracy improves in complex real life oriented datasets which have speakers from all age groups i.e. children to seniors (4–84 years). Our results also shows how gender recognition accuracy improves within particular age groups, through the multi-task learning algorithm. The proposed multi-task learning DNN uses raw speech waveforms as input, which gives the learning process freedom to learn the gender discriminative features within the DNN. Throughout the experiments, we observed that raw waveform based feature learning improve gender recognition accuracies compared to that of 40-MFCC based DNN. We also compare two different approaches of modeling speech temporal dynamics namely temporal convolution in the form of TDNN layers and LSTM based recurrent projection layers. Experimental results shows that LSTM based recurrent projection layers based DNN provides better accuracy for the proposed problem. We also provide a detail comparative analysis of single-task learning DNN with only gender target and multi-task learning DNN with both gender and age targets. Such analysis shows the effectiveness of using age as an additional target to the DNN. Further, we compare single-task learning DNN based approach with GMM based approach and prove the novelty of the proposed DNN based approach over GMM. The DNN based approach of using age an additional element is expected to improve performance of gender recognition systems deployed as part of intelligent HMI systems.

References

Alhussein, M., Ali, Z., Imran, M., & Abdul, W. (2016). Automatic gender detection based on characteristics of vocal folds for mobile healthcare system, Mobile Information Systems, vol 2016, 1-12.

Cheng, G., Peddinti, V., Povey, D., Manohar, V., Khudanpur, S., Yan, Y. (2017). *An exploration of dropout with LSTMs, in Proceedings of Interspeech 2017, The 11th Annual Conference of the International Speech Communication Association*, August 20–24, Stockholm, Sweden.

Ghahremani, P., Manohar, V., Povey, D., & Khudanpur, S. (2016). *Acoustic modelling from the signal domain using CNNs, In Proceedings of Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, September 8–12, San Francisco, CA, USA.

Goel, N. K., Sarma, M., Kushwah, T. S., Agrawal, D. K., Iqbal, Z., & Chauhan, S. (2018). *Extracting speaker's gender, accent, age and emotional state from speech, In Proceedings of Interspeech 2018, The 19th Annual Conference of the International Speech Communication Association*, (pp. 2–6). Hyderabad, India.

Golik, P., Tuske, Z., Schluter, R., & Ney, H. (2015). *Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR, In Proceedings of Interspeech 2015, The 16th Annual Conference of the International Speech Communication Association*, (pp. 26–30). September 6–10, Dresden, Germany.

Hebbar, R., Somandepalli, K., & Narayanan, S. (2018). *Improving Gender Identification in Movie Audio using Cross-Domain Data In Proceedings of Interspeech 2018, The 19th Annual Conference of the International Speech Communication Association*, (pp. 2–6). Hyderabad, India.

Hermansky, H., & Sharma, S. (1998). *TRAP- Classifiers for Temporal Patterns, In Proceedings of 5th International Conference On Spoken Language Processing*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Jaitly, N., & Hinton, G. (2011). *Learning a better representation of speech sound waves using restricted Boltzmann machines, In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 5884–5887). Prague, Czech Republic.

Kabil, S. H., Muckenhirn, H., & Doss, M. M. (2018). *On Learning to Identify Genders from Raw Speech Signal using CNNs, In Proceedings of Interspeech 2018, The 19th Annual Conference of the International Speech Communication Association*, (pp. 2–6). Hyderabad, India.

Kumar, N., Nasir, M., Georgiou, P., & Narayanan, S. S. (2016). *Robust multichannel gender classification from speech in movie audio, In Proceedings of Interspeech 2016, The 17th Annual Conference of the International Speech Communication Association*, (pp. 8–12). San Francisco, USA.

Levitan, S. I., Mishra, T., & Bangalore, S. (2016). *Automatic Identification of Gender from Speech, In Proceedings of Interspeech 2016, The 17th Annual Conference of the International Speech Communication Association*, (pp. 8–12). San Francisco, USA.

Li, M., Han, K. J., & Narayanan, S. (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech and Language, 27*, 151–167.

Li, M., Jung, C., & Han, K. J. (2010). *Combining five acoustic level modeling methods for automatic speaker age and gender recognition, In Proceedings of Interspeech 2010, The 11th Annual Conference of the International Speech Communication Association*, (pp. 26–30). Makuhari, Chiba, Japan.

Meinedo, H., & Trancoso, I. (2010). *Age and gender classification using fusion of acoustic and prosodic features, In Proceedings of Interspeech 2010, The 11th Annual Conference of the International Speech Communication Association*, (pp. 26–30). Makuhari, Chiba, Japan.

Meinedo, H., Trancoso, I., & (2011). Age and gender detection in the I-DASH project. *ACM Transactions on Speech and Language Processing*, *7*(4), 13.

Palaz, D., Magimai-Doss, M., & Collobert, R. (2015). *Analysis of CNN-based speech recognition system using raw speech as input, In Proceedings of Interspeech, The 16th Annual Conference of the International Speech Communication Association*, September 6–10, Dresden, Germany.

Palaz, D., Magimai-Doss, M., & Collobert, R. (2019). End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. *Speech Communication*, *108*, 15–32.

Peddinti, V., Povey, D., & Khudanpur, S. (2015). *A time delay neural network architecture for efficient modeling of long temporal contexts, In Proceedings of Interspeech 2015, the 16th Annual Conference of the International Speech Communication Association*, September 6–10. Dresden, Germany.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). *The kaldi speech recognition toolkit, In Proceedings of IEEE 2011 Workshop on*

Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US.

Povey, D., Zhang, X., & Khudanpur, S. (2015). *Parallel training of deep neural networks with natural gradient and parameter averaging, In Proceedings of International Conference on Learning Representations (ICLR)*.

Ruder, S. An overview of multi-task learning in deep neural networks. Retrieved from https://arxiv.org/pdf/1706.05098.pdf. Accessed 25 Feb 2018.

Sainath, T. N., Kingsbury, B., Mohamed, A., & Ramabhadran, B. (2013). *Learning Filter Banks within a Deep Neural Network Framework, In Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 297–302.

Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W., & Vinyals, O. (2015). *Learning the speech front-end with raw waveform CLDNNs, In Proceedings of Interspeech, The 16th Annual Conference of the International Speech Communication Association*, September 6–10, Dresden, Germany.

Sak, H., Senior, A., & Beaufays, F. (2014). *Long short–term memory recurrent neural network architectures for large scale acoustic modeling, In Proceedings of the Interspeech 2014-15th Annual Conference of the International Speech Communication Association*, September 14–18, Singapore.

Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., & Dehak, N. (2018). *Emotion Identification from raw speech signals using DNNs, In Proceedings of Interspeech 2018, The 19th Annual Conference of the International Speech Communication Association*, (pp. 3097–3101). Hyderabad, India.

Speaker Recognition Evaluation. (2000). Retrieved from https://catalog.ldc.upenn.edu/LDC2001S97. Accessed 3 Nov 2018

Shobaki, K., Hosom, J., & Cole, R. A. (2000).*The OGI kids' speech corpus and recognizers, In Proceedings of Interspeech 2000, The 6th International Conference on Spoken Language Processing, ICSLP 2000 / Interspeech 2000*, (pp. 16–20). Beijing, China.

Switchbord Cellular Part-I. Retrieved from https://catalog.ldc.upenn.edu/LDC2001S13. Accessed 18 Feb 2019.

The 2008 NIST Speaker Recognition Evaluation Results. Retrieved from https://www.nist.gov/itl/iad/mig/2008-nist-speaker-recognition-evaluation-results. Accessed 15 Mar 2017.

The NIST Year 2010 Speaker Recognition Evaluation Plan. Retrieved from https://www.nist.gov/system/files/documents/itl/iad/mig/NIST_SRE10_evalplan-r6.pdf. Accessed 30 Mar 2017

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., Zafeiriou, S. (2016). *Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016*, March 20–25, Shanghai, China.

Tuske, Z., Golik, P., Schluter, R., & Ney, H. (2014). *Acoustic modeling with deep neural networks using raw time signal for LVCSR, In Proceedings of Interspeech 2014, The 15th Annual Conference of the International Speech Communication Association*, (pp. 890–894). 14–18 Singapore.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(3), 328–339.