A DEEP NEURAL NETWORK BASED END TO END MODEL FOR JOINT HEIGHT AND AGE ESTIMATION FROM SHORT DURATION SPEECH

Shareef Babu Kalluri[1], *Deepu Vijayasenan*[1], *Sriram Ganapathy*[2]

[1]Department of E & C Engineering,
National Institute of Technology Karnataka-Surathkal, Mangalore, India
[2]Learning and Extraction of Acoustic Patterns, (LEAP) Lab,
Electrical Engineering, Indian Institute of Science, Bengaluru, India

ABSTRACT

Automatic height and age prediction of a speaker has a wide variety of applications in speaker profiling, forensics etc. Often in such applications only a few seconds of speech data is available to reliably estimate the speaker parameters. Traditionally, age and height were predicted separately using different estimation algorithms. In this work, we propose a unified DNN architecture to predict both height and age of a speaker for short durations of speech. A novel initialization scheme for the deep neural architecture is introduced, that avoids the requirement for a large training dataset. We evaluate the system on TIMIT dataset where the mean duration of speech segments is around 2.5s. The DNN system is able to improve the age RMSE by at least 0.6 years as compared to a conventional support vector regression system trained on Gaussian Mixture Model mean supervectors. The system achieves an RMSE error of 6.85 and 6.29 cm for male and female height prediction. In case of age estimation, the RMSE errors are 7.60 and 8.63 years for male and female respectively. Analysis of shorter speech segments reveals that even with 1 second speech input the performance degradation is at most 3% compared to the full duration speech files.

Index Terms— Automatic Joint Height and Age Estimation, Support Vector Regression, Deep neural network, Short duration.

1. INTRODUCTION

Speech contains information about linguistic content as well as speaker identity and paralinguistic information like age, height, gender and emotional state. Estimating the physical parameters like height and age of a speaker helps in applications like forensics and commercial scenarios. For example, in voice surveillance applications, predicting the speaker meta data from the short chunks of speech data is crucial for biometric evidence generation. Similarly, predicting age and gender of a speaker from the speech data helps targeted advertisements based on the speaker profile (youth, adults and late adults) [1]. In addition, the speaker profiling methods can aid speaker diarization and verification applications as well.

In this work, we aim to predict the age and height of a speaker from short duration speech inputs ($1-3$s). We propose a DNN architecture to predict speaker height and age jointly. We explore a novel scheme to initialize the network using a conventional system based on Support Vector Regression (SVR) trained with Gaussian Mixture Model-Universal Background Model (GMM-UBM) supervector features. This initialization eliminates the need for large amounts of data for the deep neural network training. To the best of our knowledge, this is the first attempt to develop an end-to-end model that predicts the height and age of a speaker jointly.

The organization of the paper is as follows. In Section 2, we provide details of the previous attempts to height and age prediction. Section 3 describes the baseline system [2]. Section 4 details the proposed deep neural network architecture. Section 5 reports the dataset used, DNN model initialization method, and the results of various height/age prediction experiments. Finally, Section 6 concludes the paper.

2. RELEVANT PRIOR WORK

In the past, several researchers have made partially successful attempts in predicting the height and age of a speaker independently from the speech data. The main motivation for height prediction from speech comes from the scientific studies of magnetic resonance imaging by Fitch and Giedd which shows that the vocal tract length is correlated with the individual's height [3]. Thus, the automatic height prediction methods attempt to infer the vocal tract length from speech which in turn predicts the height of the speaker. A number of features have been proposed for predicting the speaker height. The widely used toolkit for extracting a range of these features is the *Open-Smile* toolkit, which extracts the statistics (mean, median, percentiles etc.) of the short-term spectral features [4, 5]. Another feature set that is extracted for this application uses the sub-glottal resonance frequencies to predict the height of a speaker [6]. These resonance frequencies are shown to be correlated to the speaker height information, and a simple, direct relation is employed to predict the height. Subsequently linear support vector regression is performed to obtain the predicted value.

Variation of age affects speech characteristics like fundamental frequency, sound pressure level, speech rate etc [7]. Also, the age of a speaker impacts the speech characteristics like jitter/shimmer [8] and speech harmonics [9]. Earlier works focused on classifying speakers into different age groups and treat this as a classification problem. Typically, Gaussian mixture model universal background model (GMM-UBM) is used for this task. The means of mixture components obtained from the GMM are represented as a supervector / i-vector for each utterance. These are used to train a support vector machine (SVM) to classify the age group (child, young adult, adult and late adult) [10, 11]. Later, an i-vector based approach for age prediction from long telephone conversational speech was proposed [12]. This uses a support vector regression (SVR) for age prediction from the i-vector. A similar algorithm was developed using long short-term memory (LSTM) networks instead of the SVR [13].

This work is partially supported by Science and Engineering Research Board (SERB) under grant no: EMR/2016/007934.

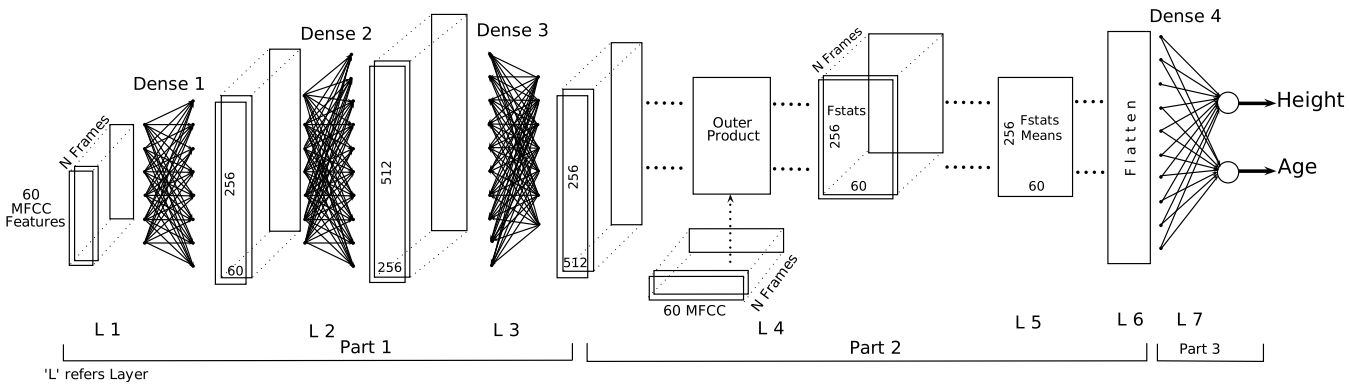

Fig. 1. Block diagram of deep neural network architecture for joint prediction of height and age of a speaker from speech

Recently, an x-vector framework was proposed as a substitute for i-vectors in speaker recognition [14]. Ghahremani *et al.* proposed an end-to-end deep neural network (DNN) using the x-vectors for age prediction [15]. However, the shortest duration speech input considered for many of the previous approaches is in the range of $5-10$ seconds. This may be too long for many forensics / profiling scenarios. Furthermore, the training of i-vector / x-vector methods are data intensive. Hence, most of the prior research efforts were trained on large datasets like NIST-2008 SRE [16].

To our best knowledge, the only work to use a common set of features for height and age predictions from short duration speech inputs is by Singh *et al.* [17]. This work attempted to predict the age and height of a speaker from a Bag of Words representation of short-term spectral features with different time resolution.

3. BASELINE SYSTEM

Our baseline system is trained with linear support vector regression model using first order statistics computed from a GMM model. We train a GMM-UBM with diagonal covariance using cepstral features of the train data. For a given the sequence of input feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$, the density function of GMM is given by,

$$p(\mathbf{x}) = \sum_{k=1}^{M} w_k \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_k, \mathbf{C}_k), \tag{1}$$

where $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$, $\boldsymbol{\mu}_k$ denote the input feature vector and mean respectively and $\boldsymbol{C}_k$ represents diagonal covariance matrix of the k^{th} GMM component with weight w_k. The frame level first order statistics (defined as $\mathbf{f}_i^j$ for a given frame i is computed as,

$$\mathbf{f}_i^j = \mathbf{x}_i p(j|\mathbf{x}_i), \tag{2}$$

where the *a-posterior* probabilities $p(j|\mathbf{x}_i)$ are computed by the Bayesian rule, given as follows,

$$p(j|\mathbf{x}_i) = \frac{w_j \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}_j, \mathbf{C}_j)}{\sum_{k=1}^{M} w_k \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}_k, \mathbf{C}_k)}. \tag{3}$$

We concatenate all mixture component specific stats $\mathbf{f}_i^j$ to form a frame level super vector $\mathbf{F}_i$. We perform the mean across time to get first order statistics $\mathbf{F}$ (refered as Fstat) across the entire speech utterance.

$$\mathbf{F} = \frac{1}{T} \sum_{i=1}^{T} \mathbf{F}_i \tag{4}$$

The vector $\mathbf{F}$ (called Fstat) is used as the feature representation for the regression system [2]. Separate SVR models are trained to predict the physical parameters like height and age. As the dimension of the input features is high, we used a linear SVR. The prediction model of the linear SVR for an input frame $\mathbf{x}_i$ is given by,

$$H_i = \sum_{i=1}^{n_s} \mathbf{v}_i^T \mathbf{F} + b = \mathbf{w}^T \mathbf{F} + b \tag{5}$$

where n_s is the number of support vectors $\mathbf{v}$, b is the bias, and $w = \sum_{i=1}^{n_s} \mathbf{v}_i$. The prediction output H_i indicates the height/age estimate for the current feature vector $\mathbf{x}_i$. The average prediction (averaged over the frames $1...T$) is used as the estimate of height/age for the utterance. The SVR models are trained and evaluated separately for male and female speakers.

4. DEEP NEURAL NETWORK ARCHITECTURE

The proposed deep neural architecture for joint prediction of height and age parameters is inspired from our baseline algorithm. The block diagram of the proposed DNN model is shown in Fig. 1.

The model has three parts. The first part (Layers L1, L2, L3) corresponds to GMM posterior computation (Eqn: 3). The second part (Layers L4, L5, L6) performs statistics computation (Eqn: 4) and the final part (Layer L7) represents the SVR regression (Eqn: 5). The first part is a fully connected multilayer perceptron shared among all the input speech frames. The second part performs frame wise first order statistics computation and computes the mean along time to get the statistics across the entire speech utterance. The trainable parameters of the network are in Part 1, 3.

Typically, deep neural network (DNN) architectures generally require a lot of training data to learn the parameters. Further, the model has to be efficient to perform regression on very short duration variable length speech segments. We exploited our baseline system for an innovative approach for the initialization of the neural network.

Since we envisage the first part of the network to predict the GMM posteriors, we initialize these layers from a smaller network trained to predict the GMM posteriors of the baseline system. A three layer network is trained separately for this purpose. The network targets for training are obtained as the GMM-UBM frame level posteriors. The network has ReLU non linearities in the hidden layers and softmax at the output layer. The network parameters are

learned over the entire training data. The second part of the network exactly replicates the operations performed in Eq. 2 and Eq. 4 where the posteriors $p(j|\mathbf{x}_i)$ are obtained using the neural network (first part of the network). The third part of the network is about predicting the speaker parameters from the first order statistics. The network is trained with sum of mean square error in age and height prediction. We initialize this layer from the baseline linear SVR. The weights corresponds to height and age targets are initialized from the respective SVR models. Following the initialization, the network is trained using back propagation with a mean square error loss. We learn separate models for male and female speakers.

5. EXPERIMENTS AND RESULTS

We perform all the experiments on the TIMIT dataset. The standard train-test split is used in all experiments. We consider the male and female speakers separately. In the dataset, there were 461 speakers (326 male and 135 female) for training and validation, 162 speakers (112 male and 56 female) for testing, where each speaker contributes ten recordings. The height values of the training data range from 145cm to 199cm and testing data range from 153cm to 204cm. Similarly the age values of training data range from 21 years to 76 years and testing data range from 22 years to 68 years. There is no overlapping of recordings of speakers in train and test splits. The duration of the recordings range from $1-6$s with an average of about 2.5s.

We use the standard performance evaluation metric of root mean square error (RMSE) to evaluate the prediction error of height and age. RMSE is defined as,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_{true,i} - x_{pred,i})^2}{N}} \quad (6)$$

where x_{true} are the true values and x_{pred} are the predicted target values of each utterance i. N is the total number of utterances.

5.1. Baseline System

We extract 20 Mel Frequency Cepstral Coefficients (MFCC) along with the delta and double delta features (feature dimension 60) from windowed speech. We perform a voice activity detection and cepstral mean and variance normalization for the input MFCC coefficients. A 256 component diagonal GMM-UBM is learned from the combined training data of male and female speakers. The first order statistics for each speech utterance is computed as described in Section 3. A separate SVR for age and height for male as well as female speakers are learned from Fstats of the training data. We call this method GMM-UBM-SVR. Table 1 details the results of this algorithm as well as comparison with a state of the art algorithm on the same task [17]. The table also lists the results of the default predictor that predicts the training mean value for all test samples. It can be noted that the age prediction algorithm of Singh *et al.* [17] is only marginally better than the default predictor.

5.2. DNN Model Initialization

As detailed in Section 4, the first part (Layers L 1 to L 3) is the equivalent of GMM posterior extraction from input MFCC features in the baseline system. Initially, this part is separately trained using the posteriors of GMM-UBM as the target values. The GMM posteriors are computed using Eqn. 3. The first part network has 2 hidden layers with 256, and 512 hidden neurons and 256 output neurons (corresponding to 256 component GMM). Both hidden layers have a dropout (0.3) and batch normalization operations. The network is trained using backpropagation to minimize the cross-entropy objective function on the TIMIT training data. The training data contains both male and female speakers. This initialization is common for both male and female models.

Table 1. RMSE values of baseline height and age estimation algorithms

Physical parameter	Singh *et al.*[17]		Default predictor	
	MALE	FEMALE	MALE	FEMALE
Height(cm)	6.70	6.10	7.01	6.51
Age(y)	7.80	8.90	8.07	9.15
Physical parameter	**GMM-UBM-SVR**		**DNN-postr-SVR**	
	MALE	FEMALE	MALE	FEMALE
Height(cm)	6.93	6.30	6.93	6.29
Age(y)	8.22	9.50	8.23	9.5

Table 2. RMSE values from DNN model for segment wise and complete duration prediction

Physical parameter	DNN-var-pred		DNN-seg-pred	
	MALE	FEMALE	MALE	FEMALE
Height(cm)	**6.85**	**6.29**	6.87	6.30
Age(y)	**7.60**	**8.63**	7.61	8.65

In order to check the sanity of the trained network, we use the trained DNN posteriors to compute the first order statistics and learn an SVR to predict speaker parameters. We denote the system as DNN-postr-SVR. Table 1 presents the corresponding results. It can be seen that the DNN posteriors are attaining very similar performance measures as the GMM-UBM.

The fully connected layer in the third part of the network is initialized from individual linear support vector regression algorithms. The male (female) neural network model is initialized from the male (female) SVR weights for height and age prediction.

5.3. DNN Learning

While the network supports variable length inputs for training, we trained it using fixed length speech inputs. We use the Keras toolkit [18]) for model learning. We have windowed the input speech into 1 second segments with 0.1-second shift. These short segments along with the corresponding target values are used as the training input for the neural network. The mean square error in height and age is used as the objective function. About 10% of the training data is kept as validation data. The validation performance is used as the training stopping criterion.

The trained network is used for height and age prediction of the test utterances. Note that the test utterances are variable length in nature. This scheme was denoted as DNN-var-pred. Table 2 reports the Deep neural network results. It can be seen that the RMSE error of age prediction has improved in both the cases over the DNN-postr-SVR system. The RMSE improvement in case of age prediction is around 0.6 years and 0.9 years for male and female speakers respectively. This is achieved without degrading the RMSE for height prediction.

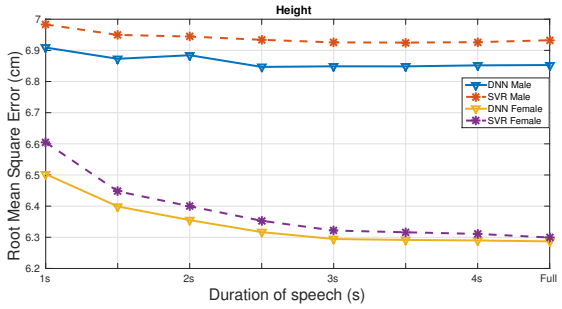

Fig. 2. RMSE of height prediction using different lengths of speech data of both male and female speakers

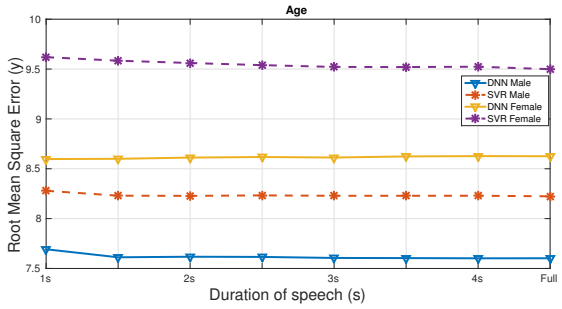

Fig. 3. RMSE of age prediction using different lengths of speech data of both male and female speakers

As a sanity check, authors have trained the model without any initialization to the DNN, the error performance is worse than the default predictor (refer Table 1).

Since the neural network is trained on 1s segments, we also tried to predict the physical parameters using windowed 1-second segments with 0.1-second shift from the variable length speech utterance. The predictions are then averaged to compute the final prediction. The result of this scheme (denoted by DNN-seg-pred) is listed in Table 2. The final RMSE values are within ± 0.05 of the DNN-var-pred scheme. Thus, even though the network was trained on 1-second length segments, it is able to generalize to variable length speech utterances.

5.4. Effect of utterance length

To analyze how shorter segments degrade the performance, we evaluated the GMM-UBM-SVR and DNN-var-pred systems with trimmed speech segments from the test data. We trim the input speech segments to different durations from $1-4$ seconds. The variation of RMSE of height prediction with respect to test speech duration is shown in Fig. 2. Even with 1 second duration, the degradation in the DNN system performance is 1.7% for male speakers and 3.2% for female speakers. The GMM-UBM-SVR system has an RMSE that is around 0.1cm more than the DNN system for 1s speech input. When the duration increases, the DNN system RMSE error improves as expected and reaches a saturation around 3s. The GMM-UBM-SVR system [2] has a higher RMSE error consistently compared to the proposed joint model.

The corresponding variations for age prediction is shown in Fig. 3. With only 1s speech available for prediction, the DNN model degrades only by 1.2% and 0.3% for male and female speakers. The RMSE of the GMM-UBM-SVR system is consistently more than the DNN system by 0.6 years for male speakers and 1 year for female speakers. Again the performance measure saturates around 3 seconds for the DNN system.

5.5. Error Analysis

In order to understand the errors, RMSE for height/age prediction is computed across different bins in the target values. Table 3 lists the results. In the training data, the height distribution is somewhat Gaussian shaped with lesser training data available for height values far away from the mean. In the results (Table 3), it can be noted that the height prediction RMSE is very high for the two extreme bins where the number of speakers are less as compared to the center bins. However, in the case of age, the training data has a more uniform distribution, and it can be observed from Table 3 that the RMSE values do not change as much as in the case of height prediction. We hypothesize that height prediction can be further improved with a more uniform training data distribution.

Table 3. RMSE values of test speakers for different bins

Height (cm)				
Range	MALE		FEMALE	
	# Train spkrs	Test	# Train spkrs	Test
$h < 150$	–	–	2	–
$150 < h < 160$	2	–	20	10.84
$160 < h < 170$	15	12.49	75	2.92
$170 < h < 180$	137	5.76	35	7.17
$180 < h < 190$	140	3.64	3	14.80
$190 < h$	32	12.98	–	–
Age (years)				
Range	MALE		FEMALE	
	# Train spkrs	Test	# Train spkrs	Test
$a < 25$	67	7.54	47	6.70
$25 < a < 30$	132	6.21	46	5.11
$30 < a < 35$	66	6.88	14	5.95
$35 < a < 40$	28	6.65	9	7.45
$40 < a < 45$	13	9.67	9	3.80
$45 < a$	20	5.98	10	8.74

6. CONCLUSIONS

In this work, we have proposed a deep neural network architecture to jointly predict speaker height and age from short duration speech segments. The neural network is initialized in a novel way using a conventional feature extraction (GMM-UBM supervectors) and regression (SVR) scheme to avoid the requirement of a large amount of data. The network is trained with mean square error criterion and the joint model is able to improve the RMSE of age prediction by more than 0.6 years, without degrading the RMSE for height prediction. Analysis of shorter durations of speech reveals that the network only degrades around 3% at most with only 1 second of the speech input. Also, the performance saturates around 3seconds. The age prediction RMSE is lower than what is reported in literature [17] that used stand-alone age prediction. In summary, the main contribution of the work is the development of a joint model (DNN) for height/age prediction which is initialized in a novel way that enables the model to perform well on short duration speech utterances.

7. REFERENCES

[1] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan, "Paralinguistics in speech and languagestate-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[2] Kalluri Shareef Babu and Deepu Vijayasenan, "Robust features for automatic estimation of physical parameters from speech," in *IEEE Region 10 Conference, TENCON 2017*, 2017, pp. 1515–1519.

[3] W Tecumseh Fitch and Jay Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[4] Todor Ganchev, Iosif Mporas, and Nikos Fakotakis, "Audio features selection for automatic height estimation from speech," in *Hellenic Conference on Artificial Intelligence*. Springer, 2010, pp. 81–90.

[5] Iosif Mporas and Todor Ganchev, "Estimation of unknown speakers height from speech," *International Journal of Speech Technology*, vol. 12, no. 4, pp. 149–160, 2009.

[6] Harish Arsikere, Steven M Lulich, and Abeer Alwan, "Estimating speaker height and subglottal resonances using MFCCs and gmms," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 159–162, 2014.

[7] Susanne Schötz, "Acoustic analysis of adult speaker age," in *Speaker Classification I*, pp. 88–107. Springer, 2007.

[8] Christian Müller and Felix Burkhardt, "Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[9] Ming Li, Kyu J Han, and Shrikanth Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[10] Mohamad Hasan Bahari, ML McLaren, DA Van Leeuwen, et al., "Age estimation from telephone speech using i-vectors," in *Proceedings of Interspeech*. 2012, Portland, USA.

[11] Amir Hossein Poorjam, Mohamad Hasan Bahari, et al., "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE),*. IEEE, 2014, pp. 7–12.

[12] Seyed Omid Sadjadi, Sriram Ganapathy, and Jason W Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5040–5044.

[13] Ruben Zazo, Phani Sankar Nidadavolu, Nanxin Chen, Joaquin Gonzalez-Rodriguez, and Najim Dehak, "Age estimation in short speech utterances based on LSTM recurrent neural networks," *IEEE Access*, vol. 6, pp. 22524–22530, 2018.

[14] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," *Proceedings of ICASSP*, 2018.

[15] Pegah Ghahremani, Phani Sankar Nidadavolu, Nanxin Chen, Jesús Villalba, Daniel Povey, Sanjeev Khudanpur, and Najim Dehak, "End-to-end deep neural network age estimation," *Proc. Interspeech 2018*, pp. 277–281, 2018.

[16] Alvin F Martin and Craig S Greenberg, "NIST 2008 speaker recognition evaluation: Performance across telephone and room microphone channels," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[17] Rita Singh, Bhiksha Raj, and James Baker, "Short-term analysis for estimating physical parameters of speakers," in *2016 4th International Conference on Biometrics and Forensics (IWBF)*. IEEE, 2016, pp. 1–6.

[18] Chollet Franois, "Keras," *https://keras.io/*.