An Investigation of Multi-Language Age Classification from Voice

Osman Büyük[1] and Levent M. Arslan[2,3]
[1]Department of Electronics and Communications Engineering, Kocaeli University, Kocaeli, Turkey
[2]Department of Electrical and Electronics Engineering, Bogazici University, Istanbul, Turkey
[3]Sestek Speech Enabled Software Technologies Incorporation, Istanbul, Turkey

Keywords: Age Classification from Voice, Multi-language, Feed-forward Deep Neural Networks, Support Vector Machines, Gaussian Mixture Models.

Abstract: In this paper, we investigate the use of deep neural networks (DNN) for a multi-language age classification task using speaker's voice. For this purpose, speech databases in two different languages are combined together to construct a multi-language database. Mel-frequency cepstral coefficients (MFCC) are extracted for each utterance. A Gaussian mixture model (GMM), a support vector machine (SVM) and a feed-forward deep neural network (DNN) systems are trained using the features. In the SVM and DNN methods, the GMM means are concatenated to obtain a GMM supervector. The supervectors are fed into the SVM and DNN for age classification. In the experiments, we observe that the multi-language training does not degrade the performance in the SVM and DNN methods when compared to the matched training where train and test languages are the same. On the other hand, the performance is degraded for the traditional GMM method. Additionally, the SVM and DNN significantly outperform the GMM in the multi-language train-test scenario. The absolute performance improvement with the SVM and DNN is approximately 12% and 7% for female and male speakers, respectively.

1 INTRODUCTION

Today's advanced voice enabled automatic systems benefit from multiple information sources such as speaker's identity, gender, emotional state and age to enhance the user's experience. Age information can be used in an interactive voice response (IVR) system to direct the client to a specific branch of the scenario or to advertise a suitable product. IVR systems may work in different regions of the world where multiple languages are spoken. Therefore, it might be convenient to develop a multi-language age identification system.

The main problem when developing a speech processing application in an under-resourced language is the lack of a manually tagged speech database. For an age identification application, it might be difficult to find large amount of data for each age-gender combination. On the other hand, speech databases from different languages can be unified in a multi-language system to construct a larger database. This might yield an additional performance improvement as well.

In Interspeech 2010, the paralinguistic challenge is organized for age/gender identification from voice (Schuller et al., 2010). For the challenge, a speech database in the German language is distributed. The database consists of recordings from 954 speakers. For this work, we constructed another age identification database in the Turkish language. The Turkish database consists of 384 speakers. The databases in the German and Turkish languages are combined together to develop a multi-language system.

Gaussian mixture models (GMM) were introduced for speaker verification (Reynolds et al., 2000) and has become the dominant classification method for years. In the GMM-UBM method (Reynolds et al., 2000), a universal background model (UBM) is trained using a large speech database collected from various different speakers. The UBM is adapted to speaker models using target speaker's speech and maximum-a-posteriori (MAP) adaptation technique. Support vector machines (SVM) have also been successfully used for speaker verification (Campbell et al., 2006). In the GMM/SVM method (Campbell et al., 2006), GMM

Büyük, O. and Arslan, L.
An Investigation of Multi-Language Age Classification from Voice.
DOI: 10.5220/0007237600850092
In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*, pages 85-92
ISBN: 978-989-758-353-7

means are concatenated to obtain a GMM supervector. Then, the GMM supervectors are used in the SVM modeling. The proposed GMM/SVM method combines the generative and discriminative powers of GMM and SVM methods.

Recently, deep neural networks (DNN) have been one of the most popular approaches for many machine learning problems especially after the introduction of efficient methods to train the networks with huge number of parameters (Hinton et al., 2006; Deng and Yu, 2013). DNNs and its variants have resulted in state-of-the-art-performance in speech processing applications such as speech recognition (Hinton et al., 2012), speaker/language recognition (Richardson et al., 2012), speech synthesis (Zen, 2015), emotion recognition (Tashe et al., 2017) and spoof detection (Zhang, 2017). The success of the DNNs mainly comes from its deep learning mechanism in which higher level features are obtained from the lower level ones and the same lower level feature contribute to many higher level features (Deng and Yu, 2013). Its deep and flexible topology makes it a suitable method for complex classification problems. Therefore, it might be the ideal choice for the multi-language age identification task, too.

Previous research on age identification mainly focused on the selection of the best classification and feature extraction methods for the task. In (Metze et al., 2007), four methods are compared; i) a parallel phone recognizer (PPR) with mel-frequency cepstral coefficients (MFCC) features ii) a dynamic Bayesian network with prosodic features iii) a system based on the linear prediction (LP) envelope of a windowed speech signal and iv) a GMM with MFCC features. In the experiments, the best performance is obtained with PPR. The PPR method performs as good as human listeners except short utterances. In (Bocklet et al., 2008), GMM/SVM is compared to GMM-UBM using MFCC features. In the experiments, GMM/SVM outperforms GMM-UBM. In (Meinedo and Trancoso, 2010), SVM and NN based systems are trained using a set of long-term features. Additionally, the authors train a NN with perceptual linear prediction (PLP) and pitch-based short-term features. Moreover, a GMM-UBM system is realized with modulation spectrogram features proposed in (Kingsbury et al., 1998). In the experiments, the best performance is obtained with the NN using PLP + pitch features. The accuracy is significantly improved with system fusion. In (Dobry et al. 2011), the GMM/SVM method is used with MFCC features. In the paper, a dimensionality reduction method is proposed as an alternative to the classical principal component analysis (PCA) and linear discriminant analysis (LDA) methods. Authors claim that the classical methods tend to eliminate some of the relevant information from features. The proposed method improved the performance compared to a baseline system in which no dimensionality reduction is performed. Additionally, model training and testing durations significantly decreased with the reduced dimensions. In (Li et al., 2013), three baseline subsystems, namely, GMM-UBM, GMM/SVM, and SVM trained with 450 long-term features are compared with four novel subsystems; i) SVM modeling of UBM weight probabilities, ii) sparse representation of UBM weight probabilities, iii) SVM modeling of GMM maximum likelihood linear regression (MLLR) matrix and iv) SVM modeling of polynomial expansion coefficients of the syllable level prosodic features. In the experiments, score level fusion of the seven subsystems outperformed all the other stand-alone systems. In (Bahari and Hamme, 2011), a weighted supervised non-negative matrix factorization (WSNMF) is used together with a general regression neural network (GRNN) for age identification. The WSNMF is trained with GMM weight supervectors. GRNN is preferred over the other NN types since it does not require an iterative training and can be used effectively for sparse data. In the experiments, the proposed method performs better than the chance level. In (Grzybowska and Kacprzak, 2016), i-vector approach is used for age identification. In the method, i-vectors corresponding to each age class are averaged in the training phase. The cosine distance between the test's and target age classes' i-vectors is computed during the test. The proposed method achieved a state-of-the-art performance on the paralinguistic challenge's German database. A DNN based approach is proposed for age identification in (Qawaqneh et al., 2017a; Abu Mallouh et al, 2017; Qawaqneh et al., 2017b). In (Qawaqneh et al., 2017b), a feed-forward DNN is used as a bottleneck feature extractor. The bottleneck DNN has 5 hidden layers with 1024 nodes in each layer except the last layer where the number of nodes is reduced to 39. As a result, it extracts the most relevant features for the classification problem. MFCCs are used as the baseline feature set in the study. The transformed MFCCs after the bottleneck DNN are fed into an i-vector and a DNN based classifiers. The transformed MFCCs achieved 13% performance improvement over the baseline MFCCs. However, the proposed method requires the use of tied-state tri-phone labels from a speech recognizer.

Previously in (Buyuk and Arslan, 2018), we propose to use a feed-forward DNN for age classification from voice. In the method, GMM supervectors are fed into the DNN similar to the GMM/SVM. The proposed method is compared to other classification methods and DNN architectures in (Buyuk and Arslan, 2018). In this paper, we investigate the use of the DNN for a multi-language age classification task. For this purpose, the German and Turkish databases are combined together in order to simulate the multi-language scenario. The multi-language DNN is trained using the combined database. We compared the proposed method with the SVM and GMM methods. In the experiments, we observed that the multi-language training does not degrade the performance in the SVM and DNN methods when compared to the matched training where the train and test languages are the same. On the other hand, it significantly improves the performance in a multi-language test scenario. Additionally, the DNN and SVM methods significantly outperform the classical GMM for the multi-language case. Both methods approximately yield 12% and 7% absolute performance improvement over the baseline GMM for female and male speakers, respectively.

The remainder of this paper is organized as follows. In Section 2, we present the German and Turkish speech databases. We describe our methodology in Section 3. The experimental results are provided in Section 4. The last section is devoted to conclusions and future work.

2 DATABASES

We use two speech databases for age identification. The databases are collected in the Turkish and German languages. We divide the speakers in the databases into three age categories. Age categories are named as young, adult and senior. Age of the young speakers is between 15 and 25. It is between 26 and 40 for the adult and over 40 for the senior speakers. Speakers in each age category are also sub-grouped according to their genders. Age classification experiments are performed gender-dependent. Further details of the databases are discussed in the following two subsections.

2.1 Turkish Database

The Turkish database consists of recordings from 384 speakers. 228 of the speakers are female and 156 of them are male. The youngest speaker in the database is 15 years old and the eldest speaker is 84 years old. The recordings are taken in a soundproof room to eliminate the background noise. They are recorded in 16 kHz, 16 bits, pulse code modulation (PCM) format and re-sampled to 8 kHz. Each speaker in the database reads 200 prompted Turkish sentences. The duration of the recording changes from 3 to 10 seconds.

The distribution of the speakers according to the age-gender categories is summarized in Table 1. In the experiments, 8 speakers from each age-gender category are set aside for testing. The remaining speakers are used in the system development. As a result of the train/test partition, 1600 test utterances for each age-gender category are used in the age classification experiments.

Table 1: Distribution of the speakers according to age-gender categories in the Turkish database.

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

2.2 German Database

The German database is first distributed in the Interspeech 2010's paralinguistic challenge (Schuller et al., 2010). The speakers in the database are asked to place calls to an IVR scenario in six separate sessions. One day break is ensured between each session to provide more variations in the speakers' voices. The calls are made indoor and outdoor environments with a mobile phone. The recordings are stored at the application server as 8 kHz, 8 bits, A-law format. They are encoded and distributed in 8 kHz, 16 bits, PCM format.

There is 47 hours of speech collected from 945 speakers in the database. The average duration of the recordings is 2.58 seconds. The speakers in the database are divided into train, development and test partitions. There are 471 speakers in the train, 299 speakers in the development and 175 speakers in the test categories. In our experiments, we use the train speakers for the system development. The development speakers are used for the testing. 1500 test utterances from each age-gender category are used in the experiments to balance the number of test utterances with the Turkish database. The test utterances are distributed uniformly among the development speakers.

The speakers in the German database are grouped into the same young (ages between 15 and

25), adult (ages between 26 and 40) and senior (ages over 40) age categories as in the Turkish database. The distribution of train and test speakers according to the age-gender categories is summarized in Table 2.

Table 2: Distribution of the speakers according to age-gender categories in the German database.

	Train		Test	
	Male	Female	Male	Female
Young	64	54	38	33
Adult	35	27	19	22
Senior	105	118	74	75
Total	204	199	131	130

3 METHODOLOGY

3.1 Feature Extraction

MFCC features are commonly used in many speech processing applications. We also make use of MFCC as a common feature type for all the classification methods. In the future, we plan to investigate other features which might be more suitable to multi-language age identification task.

13 static MFCC including the logarithm of energy are extracted for each frame of speech using 26 mel-frequency filter-bank channels. 13 delta coe□cients are appended to the static MFCC resulting in a 26 dimensional feature vector. The features are extracted for 25 milliseconds window length and 10 milliseconds skip size. Cepstral mean normalization (CMN) is applied to the features. HTK toolkit is used for the feature extraction (Young et al., 2006).

3.2 Gaussian Mixture Models (GMM)

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

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

where $N(\boldsymbol{o}_t, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is a Gaussian distribution in Equation 2;

$$N(\boldsymbol{o}_t, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} * exp\left[-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i{}^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)\right] \quad (2)$$

In the equations, $w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ are the mixture weight, mean and covariance of i^{th} mixture, respectively. $\boldsymbol{o}_t$ is the t^{th} observation vector. M is the number of mixtures in the model. If the feature vectors are assumed to be independent, the recognition score can be computed as shown in Equation 3;

$$\Lambda(\boldsymbol{O}) = \sum_{t=1}^{T} log\, P(\boldsymbol{o}_t|\lambda) \quad (3)$$

In our GMM implementation, we adapt the GMM-UBM method in (Reynolds et al., 2000) to age identification problem. First, a gender-dependent UBM is trained with 256 mixtures. 30 speakers in the German and 20 speakers in the Turkish databases for each gender are set aside for the UBM training. The UBM speakers are not used in the other stages of model development and test. The UBM speakers are distributed with respect to the number of speakers in each age-gender category. UBM is adapted to an age model using the training speech from the age category and the MAP adaptation technique. The verification scores are obtained with the age class GMM. Becars toolkit is used in the GMM implementation (Blouet et al., 2004).

3.3 Support Vector Machines (SVM)

SVM tries to find the best hyperplane that separate observations of one class from another. Assume that a feature vector is represented with a feature-label pair $(\boldsymbol{x}_t, y_t)$. Here, $\boldsymbol{x}_t$ represents the feature vector and y_t represents the class label. For a two-class classification problem, we can assume that $y_t \in (+1, -1)$. Then, SVM finds the solution for the following optimization problem in Equation 4 (Boser et al., 1992; Cortes and Vapnik, 1995);

$$\min_{w,b,\varepsilon} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l} \varepsilon_i \quad (4)$$

subject to constraints in Equations 5 and 6;

$$y_i(\boldsymbol{w}^T \emptyset(\boldsymbol{x}_i) + b) \geq 1 - \varepsilon_i \quad (5)$$

$$\varepsilon_i \geq 0 \tag{6}$$

In the equations, C > 0 is the penalty parameter of the error term. $\emptyset$ is used to map the training feature vectors into a higher dimensional space. The separating hyperplane with the maximal margin are found in that higher dimensional space. Furthermore, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \emptyset(\boldsymbol{x}_i)^T \emptyset(\boldsymbol{x}_j)$ is called the kernel function. Radial basis function (RBF) is one of the frequently used kernel functions and is described in Equation 7;

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2) \tag{7}$$

where $\gamma > 0$ is kernel parameter. This parameter is usually optimized with cross-validation.

In our SVM implementation, the mean vectors of the UBM in the GMM-UBM method are concatenated to obtain a 6656 dimensional (256 mixtures * 26 features) GMM supervector. The supervector is adapted for each utterance using MAP adaptation. Adapted GMM supervectors are used as inputs to the SVM. RBF kernel is used and kernel parameters are optimized using a five-fold cross validation. In the experiments, we use LIBSVM toolkit (Chang and Lin, 2011).

3.4 Feed-forward Deep Neural Networks (DNN)

Recently, DNN has become one of the most popular modelling approaches for many machine learning problems. Introduction of efficient methods to train networks with huge number of parameters have been the main turning point in the DNN research (Hinton et al., 2006; Deng and Yu, 2013). Additionally, the drastic increase in processing power of the computing machines has accelerated the research in deep learning. DNN achieved very good performance in many machine learning problems (Hinton et al., 2012; Richardson et al., 2012; Zen, 2015; Tashe et al., 2017; Zhang, 2017). As a result, many research institutes turned their focus on this emerging field (Taigman et al., 2014; Sainath et al., 2015).

A feed-forward DNN is a multilayer perceptron with two or more hidden layers (Deng and Yu, 2013). The neurons in the consecutive layers are fully connected. In a feed-forward DNN, the activation value of j^{th} neuron in the l^{th} layer can be computed as in Equation 8;

$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right) \tag{8}$$

In the equation, w_{jk}^l is the weight between j^{th} and k^{th} neurons in the l^{th} layer, b_j^l is the bias term for the j^{th} neuron in the l^{th} layer and σ is the activation function. Sigmoid activation is frequently used and defined as in Equation 9;

$$\sigma(x) = \frac{1}{1 - e^{-x}} \tag{9}$$

In our DNN implementation, the same GMM supervectors in the SVM method are fed into a feed-forward DNN. The feed-forward DNN has two hidden layers with sigmoid activation functions. The number of neurons in each hidden layer is 100. Mean squared error is used as the loss function. DNN implementation is performed with Keras toolkit (Chollet, 2015).

4 EXPERIMENTAL RESULTS

Age classification accuracies are provided in Table 3 and Table 4 for male and female speakers, respectively. In the tables, rows represent the test and columns represent the train language. Multi-language refers to the multi-language train/test scenario. In the Turkish and German train/test cases, only the utterances from the target language are used. In the multi-language train/test cases, the train/test utterances are combined together.

We can make several observations from Table 3 and Table 4. First of all, age classification performances of all three methods are equal to chance level (chance level is 33% for a three class classification problem) for language mismatch conditions. This result might be attributed to the different acoustic spaces in the target languages. Additionally, the diverse differences between the recording conditions of the two databases might also lead to poor classification accuracy. As mentioned in Section 2, the Turkish database is collected in a soundproof room with a high quality microphone. On the other hand, the German database is collected over a telephone line in indoor and outdoor environments and includes noisy speech.

Secondly, the recognition accuracy in Turkish female speakers is much higher when compared to all other test cases. This is due to the clean recording conditions in the Turkish database. However, the recognition accuracy in Turkish male speakers is comparable to German male speakers in spite of better recording conditions. When we analyze the results in Turkish male speakers in detail, we observe that the low performance is mainly because

Table 3: Age classification accuracies (in %) for male speakers. Rows and columns represent the test and train languages, respectively.

	Turkish			German			Multi-language		
	GMM	SVM	DNN	GMM	SVM	DNN	GMM	SVM	DNN
Turkish	49.2	49.3	46.6	37.6	33.5	33.5	46.7	50.0	47.7
German	34.5	30.1	34.9	45.0	46.2	46.5	35.1	49.0	48.5
Multi-language	42.3	40.4	41.1	41.1	39.6	39.6	41.2	49.7	48.1

Table 4: Age classification accuracies (in %) for female speakers. Rows and columns represent the test and train languages, respectively.

	Turkish			German			Multi-language		
	GMM	SVM	DNN	GMM	SVM	DNN	GMM	SVM	DNN
Turkish	69.6	70.8	71.0	27.6	33.3	33.3	54.5	69.2	69.6
German	32.1	32.1	31.0	42.4	50.2	45.6	40.11	49.7	47.4
Multi-language	51.8	51.9	52.0	34.6	41.3	39.2	47.6	59.9	59.1

of the senior age class. As given in Table 1, there are only 16 senior male speakers in the Turkish database. We think that having relatively small number of speakers in the age category degrades the recognition accuracy. This result also shows the importance of the speech data from various different speakers for each category to achieve a good classification performance.

When we use DNN, the recognition accuracies in German database are 46.5% and 45.6% for the matched and 48.5% and 47.4% for the multi-language training for male and female speakers, respectively. The same classification accuracies are 46.2% and 49.0% for the matched and 50.2% and 49.7% for the multi-language training in the SVM method. These accuracies are comparable to the results reported in (Meinedo and Trancoso, 2010).

The DNN and SVM methods perform slightly better than the GMM for the language matched cases. The performances of the DNN and SVM are comparable. On the other hand, the performance gain of the DNN and SVM is significant when the training is performed in multi-language. In the multi-language test case, the absolute performance improvement of the DNN over the baseline GMM is 6.9% and 11.5% for male and female speakers, respectively. The absolute performance improvement of the SVM over the baseline GMM is 8.5% and 12.3% for male and female speakers, respectively. We think that the discriminative training of the GMM supervectors in the SVM and DNN methods results in a significant performance improvement in the multi-language test scenario. Moreover in the DNN and SVM methods, the multi-language training does not substantially degrade the performance compared to the matched training though it also does not improve it significantly. On the other hand, the multi-language training degrades the performance for all test cases in the classical GMM method.

5 CONCLUSIONS

In this study, we propose to use the GMM supervectors as inputs to a feed-forward DNN for age classification from voice. We test the performance of the proposed method using a multi-language database. The database consists of recordings from the Turkish and German languages. In the experiments, we observed that the DNN method significantly outperforms the traditional GMM for the multi-language scenario. Its performance is comparable to the SVM based method. Moreover, the performance in the DNN is not degraded substantially with the multi-language

training compared to the matched training when the test utterances are solely from one language.

The main limitation of our work is the differences in the recording conditions of the two databases. As a result, the classification results not only contain the language mismatch but also the mismatch between the recording conditions. In the future, we plan to extend our speech database with new languages recorded in more unified environmental and background conditions.

In the experiments, we observed that all three methods do not perform well in language and background mismatch conditions with MFCC features. In order to improve the performance in the mismatch conditions, we will investigate the other features which are more suitable for the multi-language age identification task. We also plan to implement deep neural network types such as recurrent neural network (RNN) and convolutional neural network (CNN) for age identification in the future.

ACKNOWLEDGEMENTS

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

REFERENCES

Abu Mallouh A., Qawaqneh Z. and Barkana B., 2017. "Combining two different DNN architectures for classifying speaker's age and gender," International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 4: BIOSIGNALS, (BIOSTEC 2017).

Bahari, M.H., and Hamme, H.V., 2011. "Speaker age estimation and gender detection based on supervised non-negative matrix factorization," IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS 2011), Milan, Italy.

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

Bocklet, T., Maier, A., Bauer, J.G., Burkhardt, F., and Noth, E., 2008. "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), Las Vegas, USA.

Boser, B.E., Guyon, I., and Vapnik, V., 1992. "A training algorithm for optimal margin classifiers," ACM Workshop on Computational Learning Theory Pittsburgh, USA. pp. 144–152.

Buyuk, O., and Arslan, L.M., 201. "Combination of long-term and short-term features for age identification from voice," Advances in Electrical and Computer Engineering 18 (2), pp. 101-108.

Campbell, W.M., Sturim, D.E., and Reynolds, D.A., 2006. "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters 13 (5), pp. 308-311.

Chang, C.C., and Lin, C.J., 2011. "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology 2 (3), pp. 27:1-27:27.

Chollet, F., 2015. "Keras," Github repository 2015. https://github.com/fchollet/keras.

Cortes, C., and Vapnik, V., 1995. "Support-vector networks," Machine Learning 20 (3), pp. 273–297.

Deng, L., and Yu, D., 2013. "Deep learning methods and applications," Foundations and Trends in Signal Processing 7 (3-4), pp. 197-387.

Dobry, G., Hecht, R.M., Avigali M., and Zigel, Y., 2011. "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," IEEE Transactions on Audio, Speech, and Language Processing 19 (7), pp. 1975-1985.

Grzybowska, J., and Kacprzak, S., 2016. "Speaker age classification and regression using i-vectors." International Conference on Spoken Language Processing (INTERSPEECH 2016), San Francisco, California, USA.

Hinton, G.E., Osindero, S., and Teh, Y., 2006. "A fast learning algorithm for deep belief nets," Neural Computation 18, pp. 1527-1554.

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

Kingsbury, B.E., Morgan, N., and Greenberg, S., 1998. "Robust speech recognition using the modulation spectrogram," Speech Communications 25, pp. 117-132.

Li, M., Han, K.J., and Narayanan, S., 2013. "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," Computer Speech and Language, 27 (1), pp. 151-167.

Meinedo, H., and Trancoso, I., 2010. "Age and gender classification using fusion of acoustic and prosodic features," International Conference on Spoken Language Processing (INTERSPEECH 2010), Makuhari, Japan.

Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B., Bauer, J.G., and Little, B., 2007. "Comparison of four approaches to age and gender recognition for telephone applications," IEEE International

Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), Hawaii, Honolulu, USA.

Qawaqneh Z., Abu Mallouh A. and Barkana B., 2017a. "DNN-based models for speaker age and gender classification," International Joint Conference on Biomedical Engineering Systems and Technologies, vol. 4: BIOSIGNALS, (BIOSTEC 2017).

Qawaqneh, Z., Abu Mallouh, A., and Barkana, B., 2017b "Deep neural network framework and transformed MFCCs for speaker's age and gender classification," Knowledge Based Systems 115, pp. 5-14.

Reynolds, D.A, Quatieri, T.F., and Dunn, R.B., 2000. "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing 10 (1–3), pp. 19–41.

Richardson, F., Reynolds, D., and Dehak, N., 2012. "Deep neural network approaches to speaker and language recognition," IEEE Signal Processing Letters 22 (10), pp. 1671-1675.

Sainath, T.N., Vinyals, O., Senior, A., and Sak, H., 2015. "Convolutional, long short-term memory, fully connected deep neural networks," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015), Brisbane, Australia.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Mueller, C., and Narayanan, S., 2010. "The Interspeech 2010 paralinguistic challenge," International Conference on Spoken Language Processing (INTERSPEECH 2010), Makuhari, Japan.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L., 2014. "DeepFace: Closing the gap to human-level performance in face verification," IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), Columbus, USA.

Tashe, I.J., Wang, Z.Q., and Godin, K., 2017. "Speech Emotion Recognition Based on Gaussian Mixture Models and Deep Neural Networks", Information Theory and Applications Workshop (ITA 2017).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., 2006. "The HTK Book (for HTK Version 3.4)," Cambridge, Cambridge University Engineering Department.

Zen, H., 2015. "Acoustic modeling for speech synthesis: from HMM to RNN," Automatic Speech Recognition and Understanding Workshop (ASRU 2015), Scottsdale, Arizona, U.S.A.

Zhang, C., Yu, C., and Hansen, J.H.L., 2017. "An Investigation of deep learning frameworks for speaker verification anti-spoofing," IEEE Journal of Selected Topics in Signal Processing 11 (4), pp. 684-694.