Combining Two Different DNN Architectures for Classifying Speaker's Age and Gender

Arafat Abu Mallouh[1], Zakariya Qawaqneh[1] and Buket D. Barkana[2]
[1]*Computer Science and Engineering Department, University of Bridgeport, Bridgeport, CT, 06604, U.S.A.*
[2]*Electrical Engineering Department, University of Bridgeport, Bridgeport, CT, 06604, U.S.A.*

Keywords: Deep Neural Network, GMM-UBM, I-Vector, Speaker Age and Gender Classification, Fine-Tuning.

Abstract: Speakers' age and gender classification is one of the most challenging problems in the field of speech processing. Recently, remarkable developments have been achieved in the neural network field, nowadays, deep neural network (DNN) is considered one of the state-of-art classifiers which have been successful in many speech applications. Motivated by DNN success, we jointly fine-tune two different DNNs to classify the speaker's age and gender. The first DNN is trained to classify the speaker gender, while the second DNN is trained to classify the age of the speaker. Then, the two pre-trained DNNs are reused to tune a third DNN (AGender-Tuning) which can classify the age and gender of the speaker together. The results show an improvement in term of accuracy for the proposed work compared with the I-Vector and the GMM-UBM as baseline systems. Also, the performance of the proposed work is compared with other published works on a publicly available database.

1 INTRODUCTION

Computerized systems such as language learning, phone ads, crime detection, and health monitoring are rapidly increasing, and this creates an urgent need for better performance. These systems can be improved by gathering correct information about speaker's age, gender, accent, and emotional state (Nguyen et al., 2010). Age and gender recognition is the ability to recognize the age and gender information from speaker's speech. A key stage in identifying speakers' age and gender is to select and extract effective features that represent the speaker's characteristics uniquely. Then a classifier uses these features to predict the speaker's age and gender. Many feature sets have been developed and evaluated in the literature. In general, these sets can be classified into three main categories: spectral, prosodic, and glottal features. One of the spectral feature sets is the Mel frequency cepstral coefficients (MFCCs), which is widely used by many researchers. The advantage of MFCCs is the ability to model the vocal tract filter in a short time power spectrum (Li et al., 2013, Metze et al., 2007). DNNs have been used effectively for feature extraction and classification in various fields, like computer vision (Nguyen et al., 2015, Zeiler, 2013), image processing and classification (Ciregan et al., 2012, Simonyan and Zisserman, 2014), and natural language recognition (Richardson et al., 2015, Yu et al., 2010). One of the main advantages of DNN is the deep architecture that transforms rich input features to strong internal representation (Baker et al., 2009). In the past, considerable research has been carried out to improve speaker's age and gender classification, but the field still needs more work for better results.

The main contributions of this work can be summarized as the following; the usage of DNN is investigated to classify the speaker's age and gender for the public age-annotated database of german telephone speech database (aGender) using MFCCs. A new method for training two DNNs is introduced, where Age-DNN and Gender-DNN are combined into a shared output layer to produce a tuned DNN, which is able to classify a speaker age and gender simultaneously. The reminder of the paper is organized as follows: A brief literature review, methodology, experimental results and discussion, and finally, a conclusion of the proposed work.

Abu Mallouh A., Qawaqneh Z. and Barkana B.
Combining Two Different DNN Architectures for Classifying SpeakerâĂŹs Age and Gender.
DOI: 10.5220/0006096501120117
In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)*, pages 112-117
ISBN: 978-989-758-212-7

2 LITERATURE REVIEW

Speakers' age and gender characteristics were studied early in the 1950's in terms of pitch and duration (Mysak, 1959). With the developments in computer technology, studies are mainly focusing on the classification of speakers' age and gender.

Ming-Li et al. (Li et al., 2013) utilized various acoustic and prosodic features to improve classification accuracies by combining GMM base, GMM-SVM mean super vector, GMM-SVM-MLLR super vector, GMM-SVM-TPP super vector, and SVM baseline system. The highest accuracy (52.7%) is attained when all methods are fused together. Metze et al. (Metze et al., 2007) studied different techniques for age and gender recognition, based on telephone applications. They compared the performance of their system with humans. The first technique they used was parallel phoneme recognizer (PPR), which is one of the early systems. The main core of this system is to create a PPR for each class in the age and gender database. They reported that the PPR system performs almost like human listeners, with the disadvantage of losing quality and accuracy on short utterances.

Kim et al. (Kim et al., 2007) built a home robot that classifies speakers' age and gender from speech utterances using MFCCs, Harmonic to Noise ratio, and Jitter and Shimmer feature sets. Three different classifiers were used in their work: multi-layer perceptron (MLP), GMM, and ANN, they achieved an overall accuracy of 96.57%, 90%, and 94.6% for age using the three classifiers respectively. Bahari et al. (Bahari, 2011) tried to estimates the speakers' age. In their work, they use weighted supervised non-negative matrix factorization (WSNMF) to reduce the dimension of the speaker HMM weight supervector model. They achieved an accuracy of 96% for gender recognition on a Dutch speech database. The achieved mean absolute error was 7.48 years for age recognition. Bahari et al. (Bahari, 2014) regress the speakers age by modeling each utterance using the I-Vector and then employing the SVR classifier.

The proposed system differs from other works in many aspects; in our system, we utilize the combination of different DNN architectures to extract the features and to classify the groups. Moreover, we introduce the idea of training two DNNs one for age and another for gender, then combining the two DNNs into one DNN that can classify speaker's age and gender.

3 METHODOLGY

The supervised training in DNNs aims to learn the optimal weights that will make the DNN classification process accurate with minimal overfitting. In this work, the supervised learning is divided into three parts; a DNN that learns the speakers age, a DNN that learns the speakers gender, and AGender-Tuning DNN that learns speakers age and gender together.

3.1 Gender-DNN

This network is dedicated to capturing the gender of each speaker. As shown in figure 1 the input for this network is the MFCCs set, and the output labels are male and female. The number of hidden layers is 5, where the number of nodes in each layer are 1024 nodes. Extracting speaker gender is easier than extracting the age or age and gender of the speaker. The achieved accuracy of Gender-DNN is expected to reach high scores, and this will make the Gender-DNN participation in other DNN networks effective.

3.2 Age-DNN

This network will learn the speaker's age, where the input is the MFCCs feature set, and the output labels are children, young, mature, and senior. As shown in figure 2, the number of hidden layers is five each consists of 1024 nodes. Decreasing the number of labels helps the classifier to achieve better results, we separated the gender labels from age labels to enable the classifier to focus on age prediction. In speech processing, it is known that age classification is harder than gender classification, we will train Age-DNN to focus and learn as much as possible about speakers age, then Age-DNN will be involved in a third DNN that utilize it.

3.3 AGender-Tuning System

In classification problems, two or more methods can be combined and utilized by fusing their results on the score level, but in these cases, the fusion may not utilize the full ability of each network. In this paper, we propose an alternative way to combine two or more networks by fine-tuning their last hidden layers' outputs.

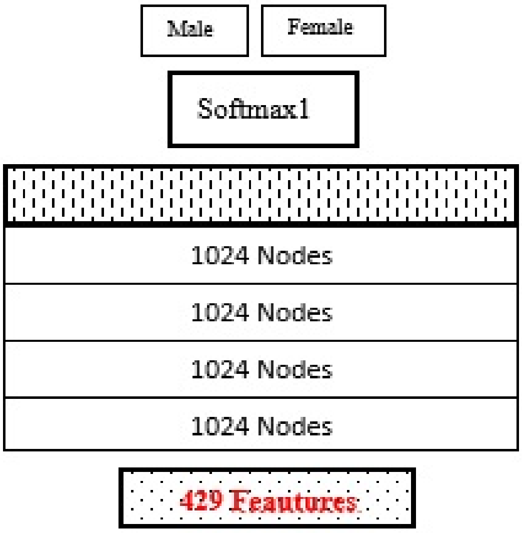

Figure 1: Gender-DNN architecture.

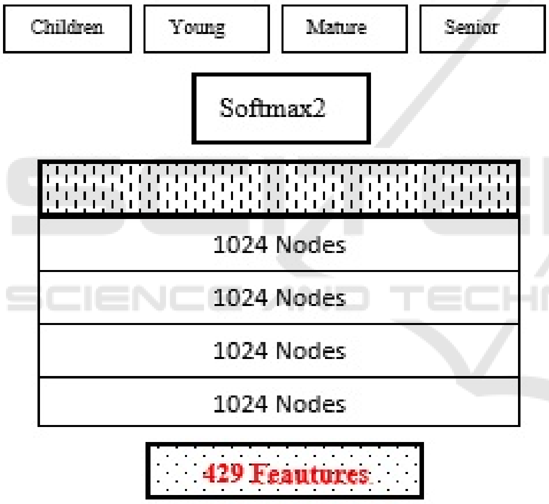

Figure 2: Age-DNN architecture.

Before combining, each network will be trained separately to utilize the network maximum ability.

First, to generate the new proposed AGender-Tuning network, the two trained age and gender networks are reused as shown in figure 3. Next, a new output layer with a softmax activation function (softmax3) is added above the last hidden layers of both networks to jointly fine tune them together. The input for the newly added output layer is the element-wise summation of the last hidden layer outputs of age network (O_1) and gender network (O_2) as in (1).

$$X = O_1 \oplus O_2 \qquad (1)$$

Where $\oplus$ is the element-wise summation between each element in the two output vectors.

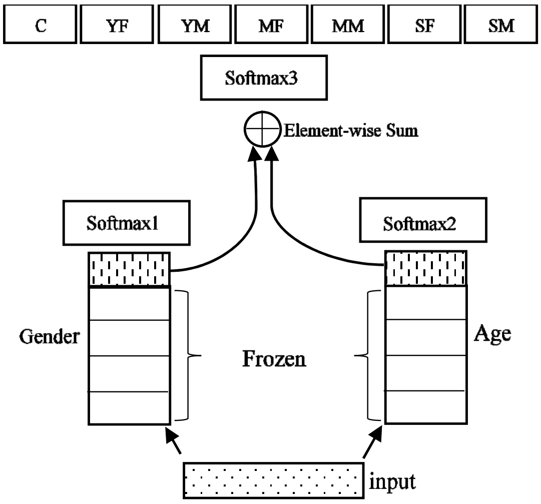

Figure 3: AGender-Tuning Network.

The output labels are 7, where each label represents a class for a group of speakers who share the same range of the age and gender. To combine the two networks, the weight values of the pre-trained age and gender networks are not changed (frozen). Then, we train and tune the weight values of the last hidden layers of the Age, Gender, and the newly added output layer as follows:

1) The newly added output layer is trained using softmax3. Consequently, the back propagation process will take effect on the last hidden layers for the Age-DNN and Gender-DNN.

2) Whenever Age-DNN receives updates from the newly added output layer, it starts updating its last hidden layer weights one more time using softmax1.

3) The same will be done for Gender-DNN, whenever Gender-DNN receives updates from the newly added output layer it; starts updating its last hidden layer weights one more time using softmax2.

4) Steps 1 to 3 are repeated until there is no learning gain.

5) Finally, after training is done, the final result (S) of speaker's age and gender classification are considered by taking the max of the newly added output layer (softmax3) as in (2).

$$S = \text{argmax } O_{\text{softmax3}} \qquad (2)$$

Where O_{softmax3} is the output posteriors of the newly added output layer (softmax3).

4 EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Data Corpus

The aGender is used to test the proposed system. The database consists of 47 hours of prompted and free text (Schuller et al., 2010). It includes seven categories: Children (C, 7-14 years old), young-female (YF, 15-24 years old), young-male (YM, 15-24 years old), mature-female (MF, 25-54 years old), mature-male (MM, 25-54 years old), senior-female (SF, 55-80 years old), and senior-male (SM, 55-80 years old). A 25% of the database of random speakers are chosen for testing, and the remaining 75% are used for training.

4.2 Baseline Systems

4.2.1 I-Vector

I-Vector is considered as one of the state-of-art systems that showed remarkable results in many fields such as speaker recognition and language identification (Richardson et al., 2015). The I-Vector process consists of, eigenvoices extraction, noise removal, and scoring. I-Vector classifier estimates different classes by using eigenvoice adaptation (Kenny et al., 2007). The process of extracting the I-Vector of any utterance starts by finding the total variability subspace. The total variability subspace is used to estimate a low-dimensional set from the adapted mean supervectors; the result is called the Identity Vector (I-Vector). For each utterance, GMM mean vectors are calculated as in (3). The UBM super vector, M is adapted by stacking the mean vectors of the GMM.

$$M = m + Tw \quad (3)$$

T represents a low-rank matrix, and w represents the required low-dimensional I-Vector. After that, the linear discriminant analysis is applied to reduce the dimension of the extracted I-Vectors by using Fisher criterion (Dehak et al., 2011). After the extraction of the I-Vectors, noise in each I-Vector is removed by Gaussian probabilistic linear discriminant analysis (Kenny, 2010). Finally, given a test utterance, the score between a target class and the test utterance is calculated using the log-likelihood ratio. In this work a 25ms MFCCs plus its delta and delta-delta are used as the input features. An age and gender independent UBM with 1024 mixtures is trained and built with a 600 dimensional I-Vector extractor, and a class subspace I-Vector of 400 dimensions for G-PLDA.

4.2.2 GMM-UBM

Seven age and gender classes were classified using a GMM classifier, with 1024 Gaussian mixtures per class. The input is 13-dimensional MFCCs with their first and second derivatives. The MFCCs are normalized to zero mean and one-unit standard deviation by using the cepstral mean subtraction and variance normalization algorithm. The UBM model is trained to inspect the general speaker’s characteristics. It is used in conjunction with the map adaptation (MAP) model by using a relevance factor of 10. In the test evaluation process, the GMM is computed for each test utterance. The log-likelihood ratio function is then used to calculate the score between each class model and the given test GMM model.

4.3 DNN Training Process

A speech utterance is divided into frames of 25 ms. In total, 39 features, one energy and 12- MFCCs with its first and second derivatives, are extracted for each frame. The number of nodes in the input layer is equal to 39×n features, where n represents the target frame concatenated with the preceding and following (n-1)/2 frames in the utterance. In the literature n is selected to be the odd numbers in {5, 21}. In our work n is chosen to be 11. The training data is divided into mini-batches, each mini-batch consists of 1024 random utterances. We used 20 epochs for training initial learning rate set to 0.1 for the first six epochs, and then decreased to one-half its original value for the remaining epochs.

4.4 Classification Accuracies

Several experiments were conducted to evaluate the performance of the proposed work and to compare it with the baseline systems. The classification accuracies are presented in Table 1. The proposed AGender-Tuning DNN outperformed the GMM-UBM system by approximately 12% and outperformed the I-Vector system by almost 7%. The proposed work extracted the speaker’s age separately from the gender before merging the last hidden posteriors of Age and Gender DNNs into one layer that is to be trained further. This separate pre-training helped to maintain the unique identity of each speaker even after age and gender posteriors became one layer.

The proposed system achieved a significant improvement especially in mature female and male classes (45.52%, 48.62%) and senior female and male classes (57.5%, 60.63%). As well as, the I-Vector classifier achieved better results than the GMM-UBM system in all classes except for the MM class. We can see that in children and SF classes, the proposed and the I-Vector systems achieved almost same results with a slight advantage for the AGender-Tuning system. The confusion matrix for the proposed system is presented in Table 2.

It can be seen from table 2 that the proposed system achieved a significant improvement for all classes especially for (MF, MM, SF, and SM). Our system was able to discriminate between these classes better than the baseline systems. Agender-Tuning system has been trained in two ways, first with separated age (Age-DNN) and gender (Gender-DNN) networks, second, with a shared output layer resulted from the Age-DNN and Gender-DNN output layers, and this shared output layer has seven age and gender labels.

Table 1: The classification accuracies of GMM-UBM, I-Vector, and AGender-Tuning DNN (%).

	GMM-UBM	I-Vector	AGender-Tuning
C	55.6	64.9	65.7
YF	48	57.1	58.3
YM	41.9	49	49.9
MF	29.6	32.5	45.5
MM	41.2	36	48.6
SF	36.4	49.9	57.5
SM	53.9	45.8	60.6

Table 2: Confusion matrix for the aGender-Tuning DNN.

Act. Pred.	C	YF	YM	MF	MM	SF	SM
C	**65.7**	14.9	5.7	3.2	2.1	6.2	2.3
YF	11.8	**58.3**	0.5	19.9	0.7	8.3	0.5
YM	2.1	0.9	**49.9**	2.6	26.9	2.9	14.7
MF	8.6	18.2	1.2	**45.5**	1.2	24.8	0.5
MM	1.2	0.1	22.3	0.3	**48.6**	1.2	26.2
SF	8.1	9.1	1.2	21.1	1.5	**57.5**	1.5
SM	1.5	0	8.7	1.2	22.2	5.8	**60.6**

To evaluate the performance of the baseline systems and the proposed work when the time duration of the speech utterance is different; we examined the overall accuracy of each system over five slots of time durations (1-5 seconds). Figure 4 shows the performance of the three systems.

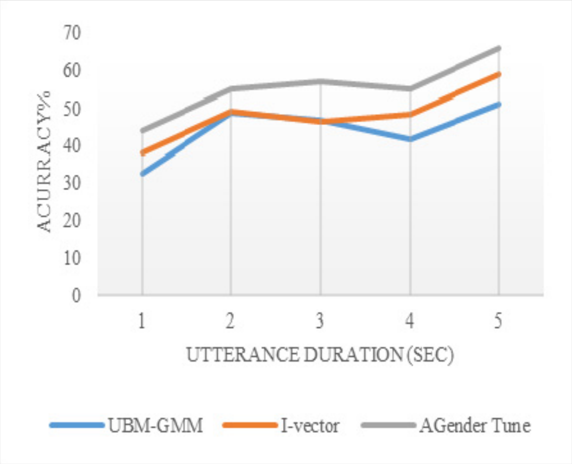

Figure 4: Comparison between the AGender-Tuning and the baseline systems for different time duration utterances.

Table 3: Comparison of proposed and previous works (%).

System	Accuracy
GMM Base-1	43.1
Mean Super Vector-2	42.6
MLLR Super Vector-3	36.2
TPP Super Vector-4	37.8
SVM Base-5	44.6
MFuse 1+2+3+4+5	52.7
AGender-Tuning	**55.16**

In general, the performance of all systems has been enhanced by increasing the duration of the utterance time; AGender-Tuning system performed better than the baseline systems for all time slots. A possible explanation refers to the fact that I-Vector and GMM-UBM systems could not build a good representation of eigenvector and GMM-UBM supervector for the corresponding utterance if it is short in time, and it is known that the aGender database utterances are short in time. When the duration of the utterance increases, for example from 3 to 4 seconds, the accuracy is not increasing for the GMM-UBM and AGender-Tuning system, this is due to the sparse data of these utterances duration, where most of the YF, MF, and MM utterances exists in this time duration. Also, we noticed from table 2 that the higher misclassification occurs between these classes and these classes have the least accuracy results among other classes.

Table 3 shows a comparison between the proposed work and six different methods presented in (Li et al., 2013). The best result in (Li et al., 2013) occurred when all systems are combined manually (MFuse 1+2+3+4+5), as shown in the table. We can see that AGender-Tuning system outperforms all the baseline systems of the manually fused system. Using our method, the accuracy of the speaker's age and gender is improved by approximately 3% compared with the fused system.

5 CONCLUSION

In this work, we proposed AGender-Tuning DNN system to classify the speakers' age and gender by combining two DNN architectures; Age-DNN to classify four groups of age, and Gender-DNN to classify the gender. A third output layer is proposed to combine the output layers of Age and Gender DNNs using element-wise summation. The results of the proposed work are compared with two baseline systems; the I-Vector and GMM-UBM on the public database aGender. The proposed work achieved better results in terms of overall accuracy and even for individual classes. Also, the proposed system was doing very well compared with the baseline systems regardless of the time duration of the speaker utterance. The overall accuracy of the proposed system, I-Vector, and GMM-UBM systems are 55.16%, 47.89%, and 43.8% respectively.

REFERENCES

Bahari, M.H. and Van Hamme, H., 2011. Speaker age estimation and gender detection based on supervised non-negative matrix factorization. In *Biometric Measurements and Systems for Security and Medical Applications (BIOMS), 2011 IEEE Workshop on* (pp. 1-6). IEEE.

Bahari, M.H., McLaren, M. and van Leeuwen, D.A., 2014. Speaker age estimation using i-vectors. *Engineering Applications of Artificial Intelligence*, *34*, pp.99-108.

Baker, J. M., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N. & Shaughnessy, D. O., 2009. Developments and directions in speech recognition and understanding, Part 1 [DSP Education]. *Signal Processing Magazine, IEEE,* 26**,** 75-80.

Ciregan, D., Meier, U. and Schmidhuber, J., 2012. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 3642-3649). IEEE.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. & Ouellet, P., 2011. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on,* 19**,** 788-798.

Kenny, P., 2010. Bayesian speaker verification with heavy-tailed Priors. *In proc. of Odyssey* - The Speaker and Language Recognition Workshop, Brno, CZ.

Kenny, P., Boulianne, G., Ouellet, P. & Dumouchel, P., 2007. Joint factor analysis versus eigenchannels in speaker recognition. *Audio, Speech, and Language Processing, IEEE Transactions on,* 15**,** 1435-1447.

Kim, H.J., Bae, K. and Yoon, H.S., 2007. Age and gender classification for a home-robot service. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*(pp. 122-126). IEEE.

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

Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B., Bauer, J.G. and Littel, B., 2007. Comparison of four approaches to age and gender recognition for telephone applications. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* (Vol. 4, pp. IV-1089). IEEE.

Mysak, E. D., 1959. Pitch and duration characteristics of older males. *Journal of Speech & Hearing Research*.

Nguyen, A., Yosinski, J. and Clune, J., 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*(pp. 427-436). IEEE.

Nguyen, P., Tran, D., Huang, X. & Sharma, D., 2010. Automatic Speech-Based Classification of Gender, Age and Accent. *In:* KANG, B.-H. & RICHARDS, D. (eds.) *Knowledge Management and Acquisition for Smart Systems and Services.* Springer Berlin Heidelberg.

Richardson, F., Reynolds, D. & Dehak, N., 2015. Deep neural network approaches to speaker and language recognition. *Signal Processing Letters, IEEE,* 22**,** 1671-1675.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C.A. and Narayanan, S.S., 2010. The INTERSPEECH 2010 paralinguistic challenge. In *InterSpeech* (Vol. 2010, pp. 2795-2798).

Simonyan, K. & Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Yu, D., Wang, S., Karam, Z. and Deng, L., 2010. Language recognition using deep-structured conditional random fields. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5030-5033). IEEE.

Zeiler, M. D., 2013. Hierarchical convolutional deep learning in computer vision. *PhD thesis, ch. 6*, New York University.