DNN-based Models for Speaker Age and Gender Classification

Zakariya Qawaqneh[1], Arafat Abu Mallouh[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, SDC, MFCCS, Speaker Age and Gender Classification.

Abstract: Automatic speaker age and gender classification is an active research field due to the continuous and rapid development of applications related to humans' life and health. In this paper, we propose a new method for speaker age and gender classification, which utilizes deep neural networks (DNNs) as feature extractor and classifier. The proposed method creates a model for each speaker. For each test speech utterance, the similarity between the test model and the speaker class models are compared. Two feature sets have been used: Mel-frequency cepstral coefficients (MFCCs) and shifted delta cepstral (SDC) coefficients. The proposed model by using the SDC feature set achieved better classification results than that of MFCCs. The experimental results showed that the proposed SDC speaker model + SDC class model outperformed all the other systems by achieving 57.21% overall classification accuracy.

1 INTRODUCTION

As the computer technology has been developing, human computer interaction (HCI) systems are becoming more important every day. Speaker age and gender information is used in some of HCI systems such as speaker identification/verification, speech recognition, tele-marketing, and security applications. Overall classification accuracies for speaker's age are quite low compared to the speaker's gender information. Until now, different feature sets have been developed and studied in the literature (Barkana, 2015). MFCCs is one of the spectral feature sets that is widely used and is able to model the vocal tract filter in a short time power spectrum (Davis, 1980). SDC feature set is reported as an effective set for language identification (LID) and speaker recognition and verification (Richardson, 2015a). SDC could be considered an extension of the delta-cepstral features, which aim to gain a significant performance over the derivative features in the cepstral features.

DNN is considered one of the most successful classifiers and feature extractors and it is used widely in different fields and applications such as computer vision (Nguyen, 2015; Zeiler, 2013), image processing and classification (Ciregan, 2012; Simonyan, 2014), and natural language recognition (Richardson, 2015b; Yu, 2010). The essential power of DNNs comes from its deep architecture which is able to transform raw input features into rich and strong internal representation (Hinton, 2012).

This paper examines and evaluates the usage of SDC feature set and DNN-based speaker model in speaker age and gender classification. Our work aims to build a model for each speaker instead of using one model for each class of speakers, whom belong to the same class of age and gender. Introducing a speaker-aware model is motivated by the fact that a speaker model can capture the characteristics of the speaker more effectively than a model that represents a group of speakers. The possible benefit of a speaker-based model is that the system can use all the features of speakers who belong to the same class to improve the classification accuracies.

The remainder of the paper is organized as follows: A brief literature review is followed by the proposed work. Then, experimental results and discussion are reported. Finally, the conclusion of our work is presented.

2 LITERATURE REVIEW

Li et al. (Li, 2013) proposed a system which combines five classifiers: Gaussian mixture model (GMM) based on MFCC features, GMM-SVM mean supervector, GMM-SVM maximum likelihood linear

Qawaqneh Z., Abu Mallouh A. and Barkana B.
DNN-based Models for Speaker Age and Gender Classification.
DOI: 10.5220/0006096401060111
In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)*, pages 106-111
ISBN: 978-989-758-212-7

regression (MLLR) supervector, GMM-SVM tandem posterior probability (TPP) supervector, and SVM baseline subsystems using 450-dimensional feature vectors including prosodic features. In addition, they combined two or more systems using fusion technique to increase the accuracy. The combination of the five systems achieved the best classification results.

Metze et al. (Metze, 2007) examined multiple classifiers for speaker age and gender classification based on telephone applications. They also compared the classification results with human performance. Four automatic classification methods, a parallel phone recognizer, dynamic Bayesian networks, linear prediction analysis, and GMM based on MFCC features are compared. Overall achieved accuracies were reported as 54%, 40%, 27%, and 42%, respectively.

Bocklet et al. (Bocklet, 2010) studied multiple systems with different combinations. A combination of several glottal, spectral, and prosodic feature sets are used in their system. They achieved an overall accuracy of 42.2% by their GMM-UBM classifier. Dobry et al. (Dobry, 2011) proposed a speech dimension reduction method for age-group classification and precise age estimation. After deploying SVM with RBF kernel, they noticed that the classifier's performance was improved by using their dimension reduction method and the SVM was faster and less affected by over-fitting problem.

Our work differs from the previous work in two ways. First, it exploits the DNN architecture to build speaker models and class models. Second, it depends on the SDC as input feature set rather than MFCCs or any other prosodic features.

Bahari et al. (Bahari, 2014) proposed i-vector model for each utterance and utilized least squares support vector regression (LSSVR) for speaker age estimation. Their work was tested on telephone conversation of national institute for standard and technology (NIST, 2010). Our work also differs from Bahari et al.'s work. While Bahari et al. attempted to regress the age of the speaker regardless of the speaker's gender, we attempt to classify the speaker age and gender at the same time. Moreover, their work depends on i-vector to represent each speaker utterance however our work uses the DNN to extract the features and to build speaker models.

3 METHODOLOGY

Typically, representing each class in age and gender classification relies on finding a general model that can capture the common characteristics of all speakers' age and gender information. In this paper, we build a model for each speaker in a class. The purpose behind this idea is to find the specific identity and concentrated characteristics of each speaker separately in order to minimize any loss of unique information related to any speaker. Since the core of this work relies on creating a model for each speaker, it is reasonable to work on a feature set that is proved to be successful in the field of speaker recognition. Motivated by the success of SDC in many speech processing fields, especially in speaker recognition, this work uses the SDC as the main feature set.

Age and gender classification problem consists of M classes, where each class has N number of speakers sharing the same age range and gender. The DNN is trained with NxM labels. The settings for the training process are given in the experimental section. After the DNN is trained, NxM speaker models are developed as shown in figure 1. The number of labels is NxM, where M is the number of classes and N is the number of speakers/class.

Each model accumulates the output layer posteriors. The accumulation of each model is done by performing feedforward on the input set until the posteriors are computed for each speaker. Then, the accumulated posteriors of the output layer are normalized (L2 normalization) and averaged for each speaker as shown in figure 2. As a result, each class will have N speaker models.

During the testing, a model will be created for the corresponding utterance using the same steps applied to build speaker models as shown in figure 2. The cosine distance is calculated between the test utterance model and every speaker model. The similarity between the test utterance and each class is computed by averaging the results of cosine similarity (Sim) between the test utterance and the speaker models belonging to the same class. Finally, the maximum similarity between the test utterance and each class is taken as the finale similarity score S as in (1).

$$S = \mathrm{Max}\left\{ \mathrm{Sim}_{cj} = \mathrm{Avg}\left(\mathrm{Sim}(c_j Spk_i, Test_{utt}) \right) \right._{i}^{N}{}_{j}^{M} \quad (1)$$

Where Avg is the statistical average function, Sim_{cj} is the cosine similarity between the test utterance and the class j, c_jSpk_i is the speaker i model of class j, and $Test_{utt}$ is the test utterance model.

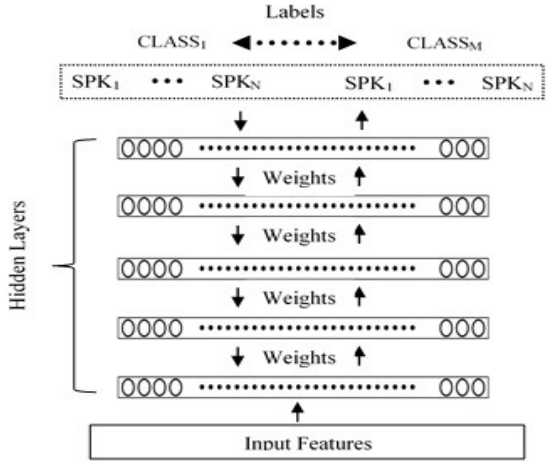

Figure 1: DNN architecture.

3.1 Score Level Fusion

The two output vectors for a given test utterance are represented as p and q and they are fused based on equation (2). p and q are the output posterior probability of SDC speaker models (SSM) and SDC class models (SCM) respectively.

$$S_j = \alpha p + (1 - \alpha) q \quad (2)$$

The final scoring for the corresponding utterance represents the index of the maximum value of the vector S_j. α is a parameter used to control the output result of the two models, and its value is set based on the performance of each model. The performance of the fusion model α values is depicted in figure 5. Several experiments are conducted to choose the optimal value of the α. The best performance occurred when α is 0.9.

3.2 Database

Age-Annotated Database of German Telephone Speech Database is used to test the proposed work. The database consists of 47 hours of prompted and free text (Schuller, 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), middle-female (MF, 25-54 years old), middle-male (MM, 25-54 years old), senior-female (SF, 55-80 years old), and senior-male (SM, 55-80 years old). One fourths of the database of random speakers is chosen for testing, and the remaining is used for training.

4 EXPERIMENTAL RESULTS

DNN architecture is used as a feature extractor and a classifier. As mentioned earlier, we evaluate the performance of our approach by using two feature sets, MFCCs and SDC. For the MFCCs, 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. 39 SDC features are extracted based on the MFCCs features as in (Campbell, 2006). The number of nodes in the input layer for both feature sets 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 an odd number between 5 and 21. In our work, n is chosen to be 11. 5 hidden layers are used, and 1024 nodes in each layer. The number of output labels equals the total number of speakers in each class. Training data is divided into mini-batches. Each mini-batch consists of 1024 random utterances. In the training process, 12 epochs are used. The learning rate is initially set to 0.1 for the first 6 epochs, and then it is decreased to one-half its initial value for the remaining epochs.

The overall classification accuracies of the MFCCs speakers models (MSM), MFCCs class models (MCM), SSM, SCM, and fused SSM+SCM are given in Table 1. The proposed SSM model achieved the best results among the other models.

The confusion matrices for the SCM, SSM, and fused SSM+SCM models are shown in Tables 2, 3, and 4. The confusion tables show that the highest misclassification rates occur between the same gender classes. In figure 3 and 4, the performance of the young (Y), middle-aged (M), and senior (S) female and male classes for all models are compared, separately. It can be seen that all models achieved somehow poor results for MF and MM classes without the score level fusion. The SSM achieved the best result for these two classes as 38.5% and 36.3%.

As shown in figure 3, for the female classes, the SSM achieved the best results except for the YF class (56%), where SCM achieved slightly better result (57.4%). This result supports the effectiveness of the SDC feature set over MFCCs. The SSM outperformed the other models in male classes (figure 4). In particular, SDC speaker and class models generated better classification results in female and male classes than MFCCs speaker and class models. However, a significant improvement (57.21%) is achieved when we fuse (SSM+SCM) models. As we can see, the fused system outperformed other models in all classes.

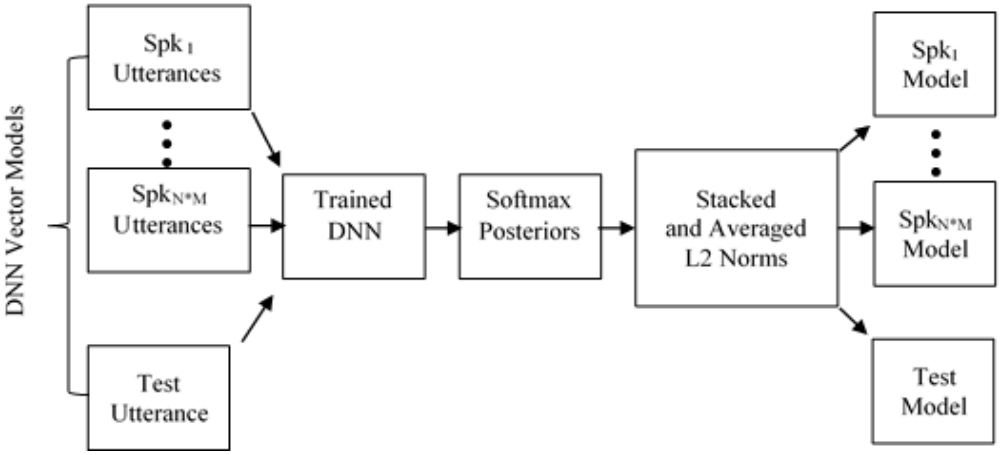

Figure 2: Flowchart of the proposed work.

Table 1: Classification accuracies (%).

	MSM	SSM	MCM	SCM	Fused (SSM+SCM)
C	56.6	58.5	57.4	60.5	74.3
YF	55.4	56	45.7	57.4	70
YM	45.1	49.9	44.3	48.3	55.4
MF	32	38.5	35.4	30.7	39.3
MM	34.3	36.3	33.8	35	39.8
SF	43.7	45.8	35.7	44.2	55.3
SM	49.3	60	49.4	57.6	66.3
%	45.2	49.3	43.1	47.7	57.2

Table 2: Confusion matrix for SCM (%).

Actual / Pred	C	YF	YM	MF	MM	SF	SM
C	60.5	17.1	8.5	3.2	3.4	6.5	0.8
YF	23.8	57.4	0.6	8.8	0.1	8.9	0.4
YM	3.3	1.8	48.3	2.4	21.0	3.2	20.0
MF	12.2	23.4	1.1	30.8	0.8	30.4	1.3
MM	1.8	0.3	27.9	1.0	35.0	3.5	30.5
SF	14.5	17.7	0.8	19.3	0.4	34.3	3.0
SM	1.0	0.3	15.6	0.9	22.1	2.4	57.7

Table 3: Confusion matrix for SSM (%).

Actual / Pred	C	YF	YM	MF	MM	SF	SM
C	58.5	18.4	8.9	3.9	3.4	5.8	1.0
YF	22.0	56.1	0.4	11.3	0.2	9.8	0.3
YM	2.3	2.1	49.9	2.3	17.3	4.7	21.4
MF	9.3	21.4	1.0	38.6	0.8	27.5	1.5
MM	1.6	0.5	26.7	1.8	36.3	3.9	29.3
SF	11.0	17.4	1.1	20.5	0.3	45.8	3.8
SM	0.6	0.2	16.9	0.9	19.3	2.0	60.1

Table 4: Confusion matrix for fused SSM+SCM (%).

Actual / Pred	C	YF	YM	MF	MM	SF	SM
C	74.3	12.9	4.3	2.6	1.3	3.3	1.4
YF	11.8	70.0	0.3	12.1	0.1	5.6	0.1
YM	1.2	0.7	55.4	1.7	19.1	3.4	18.6
MF	8.2	24.3	0.8	39.3	0.3	26.4	0.7
MM	0.5	0.0	22.3	0.4	39.8	0.4	36.6
SF	8.5	11.6	0.6	22.3	0.9	55.3	0.8
SM	1.0	0.1	9.8	0.3	19.9	2.5	66.3

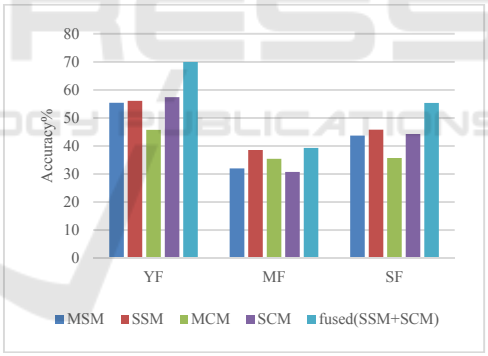

Figure 3: Classification accuracies between four methods for female speakers.

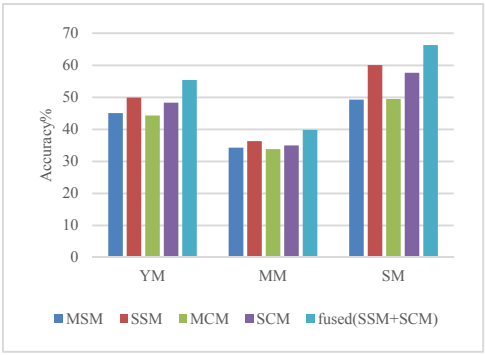

Figure 4: Classification accuracies between four methods for male speakers.

Table 5: Comparison between the proposed work and the previous works (%).

Work	System	Accuracy
(Li, 2013)	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
This work	**SDC Class Model-1**	**47.7**
	SDC Speakers Model-2	**49.3**
	Our fused model (1+2)	**57.21**

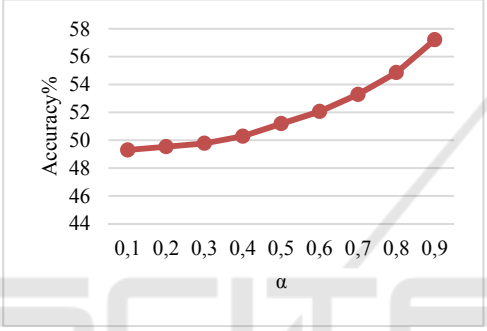

Figure 5: The performance of the fused (SSM+SCM) system versus α.

Table 5 shows the classification accuracies of the proposed models and the previous works. The best result in Li's work (Li, 2013) was achieved by fusing all the systems together manually (MFuse-1+2+3+4+5). Using our proposed models, the accuracy of the speaker age and gender classification is improved by approximately 5% when compared to the (MFuse 1+2+3+4+5). In addition, SDC class and SDC speaker models outperformed the baseline systems for the fused systems.

5 CONCLUSIONS

In this paper, we proposed DNN-based speaker models using the SDC feature set in order to improve the classification accuracies in speaker age and gender classification. The proposed speaker models and the effectiveness of the SDC feature set are compared to the class models and the MFCCs feature set as a baseline system. Our experimental results show that speaker models and the SDC feature set outperforms the class models and the MFCC set. The proposed speaker models show a better performance while classifying challenging middle-aged female and male classes where the other methods fail to classify. We compared the proposed work with the GMM Base, Mean Super Vector, MLLR Super Vector, TPP Super Vector, SVM Base, and the fused system of all these systems. The results showed that the proposed SDC speaker model + SDC class model outperformed all the other systems by achieving 57.21% overall classification accuracy.

REFERENCES

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.

Barkana, B., Zhou, J., 2015. A new pitch-range based feature set for a speaker's age and gender classification. *Applied Acoustics*, vol.98, pp.52–61.

Bocklet, T., Stemmer, G., Zeissler, V. and Nöth, E., 2010, September. Age and gender recognition based on multiple systems-early vs. late fusion. In *INTERSPEECH*, pp. 2830-2833.

Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E. and Torres-Carrasquillo, P.A., 2006. Support vector machines for speaker and language recognition. *Computer Speech & Language*, *20*(2), pp.210-229.

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.

Davis, S. and Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), pp.357-366.

Dobry, G., Hecht, R. M., Avigal, M. & 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**,** 1975-1985.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P. & Sainath, T. N., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29**,** 82-97.

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, v*ol. 4, pp. 1089-1092.

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.

NIST, The 2010 NIST Speaker Recognition Evaluation (SRE10), Link:http://www.itl.nist.gov/iad/mig/tests/sre/2010/, Accessed on 8/24/2016.

Richardson, F., Reynolds, D. and Dehak, N., 2015a. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, *22*(10), pp.1671-1675.

Richardson, F., Reynolds, D. and Dehak, N., 2015. A unified deep neural network for speaker and language recognition. *In INTERSPEECH*, vol. 2015, pp. 1146-1150.

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).

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