Improved Speaker Age Group and Gender Detection using Multiple Classifiers

Abhishek Kumar Singh
Department of Electrical Engineering
Indian Institute of Technology
Email: abhishks@iitk.ac.in

Abhay Kumar
Department of Electrical Engineering
Indian Institute of Technology
Email: abhayk@iitk.ac.in

Tanushree Gupta
Department of Electrical Engineering
Indian Institute of Technology
Email:tshree@iitk.ac.in

Ayushi Singhal
Department of Electrical Engineering
Indian Institute of Technology
Email:ayushinu@iitk.ac.in

***Abstract*—This project presents an approach to classify speakers on the basis of their age and gender. Short term features and long term features have been extracted from the voice sample of each speaker. These have been used to train Support Vector Machines and SVM Regressor. For training instances the results from these are combined to obtain the age estimate and gender of the speaker. WSNMF has been used to reduce the dimension of the short term feature space. It has performance improvement in terms of accuracy and computation time.**

***Keywords*—Age Recognition, LTF, STF, WSNMF, SVM.**

I. Introduction

In addition to its linguistic or semantic content, the acoustic speech signal also contains paralinguistic information, such as the speakers identity, accent, gender, age, or emotional state. Paralinguistic speech processing has drawn considerable interest of researchers for the last few decades [1]–[3]. In particular, gender and age recognition from speech has received much attention recently. This is mainly motivated by increased intelligent Human-Machine Interaction [4] required for various applications. In the field of speech age & gender recognition, a number of classification approaches have already been explored. Researchers tried combinations of different speech features and different classifiers for age and gender recognition.

Different possible real-life applications of speech age and gender recognition have been mentioned. Waiting queue music on phone lines can be customized according to the age and gender of the caller. This can help increase customer satisfaction of the companies. Not all people appreciate the same type of music. Older people might like slow music whereas younger people might like rock or metal music. It turns out that, a lot of times, in criminal cases the evidence is in the form of telephone speech. And by analyzing age and gender of the suspects, number of suspects can be narrowed down. Another usage of this system can be to try to understand age and gender distribution of a population in an experimental study which gives more details about the experiment.

Basically, age and gender recognition using speech consists of these three blocks :pre-signal processing, feature extraction and classification. Voice activity detection is applied to all the input speech samples to remove background noise and silences. Different short-term and long-term features have been extracted from the voiced part of the speech. For classification, various state of the art machine learning algorithms like SVM [5], Probabilistic Neural Networks.

The rest of the paper is organized as follows. Section II presents the Feature extraction from voiced part of speech. A brief description of the dimensionality reduction algorithms (WSNMF) is given in Section III. Section IV presents the details of the corpus used for the classification and training. Section V presents the details of the classification structure and the used algorithm. Section VI presents the performance evaluation and the results. A brief conclusion is presented in Section VII and Section VIII gives the scope of future work in this respect.

II. Feature Extraction

Long term (LTF) and short term features (STF) are known to provide great insight into the perceptual age and gender of the speaker. Hence LTF and STF which are supervector features (SPV) derived from MAP adaptation of means of Gaussian mixture models (GMMs) have been used to train the model. [6]

A. Long Term Features

The following long term features are extracted: pitch and the first three formants. Pitch is extracted using the autocorrelation function and the formants using Burg's algorithm. Each voice sample is windowed using a hamming window with a length of 25ms and a stepsize of 25ms, and for all the windows pitch and first three formants are calculated. The long term feature vector consists of the mean, maximum, minimum, delta, and the standard deviation over pitch and the three formants are calculated. They are concatenated to form a 20 dimensional feature vector. For a quantity x, where x could be pitch or one of the first three formants,
Mean(x) : average value of x over all the voiced speech frames.
Maximum(x): Maximum value of x over all the voiced

speech frames.
Minimum(x): Minimum value of x over all the voiced speech frames.
Delta(x) : Differnece in the maximum and minimum values of x over all voiced speech frames.
Standard deviation(x): Standard deviation of the values of x over all voiced speech frames.

B. Short Term Features

Mel Frequency Cepstral Coefficients (MFCCs) are extracted for all the utterances using a window size of 30ms, step-size of 5ms and with a dimension of 12. Each MFCC set is used to train a 128-mixture GMM model with MAP adaptation applied to constantly update the means and weights per new instance of MFCC set. The means obtained for each of the gaussians are then combined to form a 12 coefficients $\times$ 128 mixture i.e. 1536 dimensional supervector.

III. Weighted Supervised Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) is a group of algorithms in multivariate analysis where a matrix **V** is factorized into two matrices **W** and **H** where none of the matrices has any negative element. NMF has wide applications in machine learning. A popular variant of NMF used in machine learning is Weighted Supervised NMF (WSNMF)[7].

A. Weighted Supervised NMF

Consider the trainig data to be in the form $\mathbf{S^{tr}}$ where $\mathbf{S^{tr}} = (x_1, y_1), ..., (x_n, y_n), ..., (x_N, y_N)$ i.e. there are N training instances where x_n is a feature vector and y_n is a label vector consisting of 1s and 0s such that if the instance belongs to a class, the corresponding value in the label vector is 1 else it is 0.

If all elements of $\mathbf{S^{tr}}$ are non-negative then WSNMF can be applied to find a function g such that for a test instance x_{test}, y_{test} can be predicted using $y^{test} = g(x^{test})$. To this end consider

$$V_B^{tr} = [x_1\ x_2\ ...\ x_N] \tag{1}$$

$$V_S^{tr} = [y_1\ y_2\ ...\ y_N] \tag{2}$$

such that

$$V^{tr} = [V_S^{tr}; V_B^{tr}]. \tag{3}$$

NMF finds matrices W and H such that

$$V^{tr} = [W_S^{tr}; W_B^{tr}]H^{tr} \tag{4}$$

where

$$[W_S^{tr}; W_B^{tr}] = W^{tr}. \tag{5}$$

W^{tr} is obtained by minimizing the following Kullbeck-Leibler distance

$$D_{KL}(V^{tr}\|W^{tr}H^{tr}) = \sum_{mn} L_{mn}[V_{mn}^{tr} log[\frac{V_{mn}^{tr}}{(W^{tr}H^{tr})_{mn}}] + (W^{tr}H^{tr})_{mn} - V_{mn}^{tr}] + \rho \sum_{zn} H_{zn}^{tr}.$$

W^{tr} can be used to predict the labels of unseen data using

$$H^{test} = argmin_{H^{test}} D_{KL}(x^{test}\|W_B^{tr}H^{test}) \tag{6}$$

$$y^{test} = g(x^{test}) = W_S^{tr}H^{test}. \tag{7}$$

WSNMF can also be used to reduce the dimension of the input space. For this purpose, H^{tr} is used in place of V_B^{tr} and testing is performed on H^{test} instead of x^{test} and hence the reduction. WSNMF is a preferred technique because of its computational efficiency in the case of large dimension input data.

In this project WSNMF has been used to reduce the dimension of the short term supervector from 1536 to a vector of dimension 800. Finally training has been done both on the uncompressed supervector and on the dimensionally reduced vector.

IV. Corpus and Classification

We have used the Open Speech Data Corpus for German[8]. For this project voices of approximately 874 german speakers is used, each taking part in 5 sessions/utterance. The speakers are evenly distributed over the 6 target classes (see table1) with classes also being balanced for gender. The audio is recorded at 16000Hz.
For classification purpose 85% of the dataset is used for training purpose and rest for testing of the trained system. The age classes that are prevalent in our work is listed in Table1. The original age classes-schema was introduced in [german paper].

TABLE I. Classification Classes

Class index	Age group and Gender	Number of users
1	18-20,Male	146
2	21-30,Male	147
3	31-40,Male	146
4	18-20,Female	145
5	21-30,Female	146
6	31-40,Female	144

A. Classification and Regression

We divide the given data set into 6 classes. For each gender we have three age classes correspoonding to the age groups 18-20, 21-30, 31-40. Hence we have total of 6 classes (3 age classes for each gender). For the purpose of regression instead of training the regressors to predict the exact age(due to lack of specific age information) we train them to predict the catogory number.

B. Training

The training phase consist of two steps, Firstly we train SVM1 and SVM2 which directly use the LTF and the STF supervectors respectively. Further we train two gender classifiers(SVM 4 and SVM 5) and the two age regressors. For this step we use 70% of the data. Next we use 15% of the data to train SVM3 which uses the prediction results of the classifiers and the regressors trained in the first step.

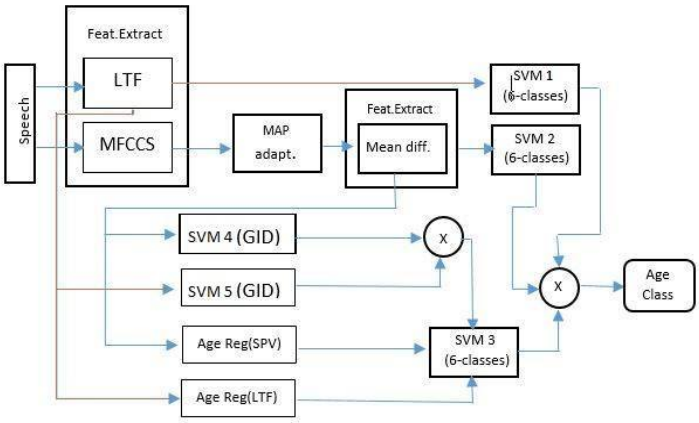

Fig. 1. Figure illustrating the classification design [6].

V. Classification Design

We use the long term features(20 features) and the short term features supervector(1536 features) to train two 6-class SVM (SVM1 and SVM2 respectively). We further train two gender classifiers(SVM) with the LTF and the STF supervectors respectively. We also train two age regressors(SVR) to predict the index of the class [6].

The posterior probabilities from the two gender classifiers and the ouputs of the two age regressors are then concatenated to form another feature vector(6 features) which is then used to train another 6-class SVM (SVM3). We also alternatively use WSNMF described in section III to compress the 1536 features of the short term feature supervector to 800 features. Final label is predicted by combining the outputs from all 3 Support Vector Machines. The detailed process is also illustrated in the Algorithm 1.

VI. Performance Evaluation

The performance of our regressors is measured in two ways: root mean squared error (RMSE), which reflects the average squared difference between the predicted and target age catogory, and correlation coefficients. By both measures, SPV-based regression RMSE/correlation values be 0.39 and 0.43 and for LTF-based regression,the RMSE /correlation values being 0.23 and 0.62 (SPV).

The two gender classifiers give accuracies of 97% (with LTF) and 88% (short term feature SPV). We see that the accuracy is higher with long term features which is possibly due to presence of pitch information in LTF which is distinctive between males and females.

The final accuracy of 6-class classification if 65.9% with long term features and 69.7% with short term feature supervector. By combining the outputs of these two classifiers the accuracy was improved to 70.45%.

If we use all three 6-class classifiers then the overall prediction accuracy can be improved to 73.5%.

Algorithm 1 : Fused Speaker Age and Gender Estimation

Input : Open Speech Data Corpus for German. Considered 5 utterences of approximately 874 speakers.
Voice Activity Detection : Detect and retain voiced part of speech. Voiced part of speech has High short-term Energy and low Zero Crossing rate.
Long Term Feature Extraction : Extract pitch using the auto-correlation function and the formants using Burgs algorithm. The long term feature vector consists of the mean, maximum, minimum, difference of maximum and minimum, and the standard deviation over pitch and the three formants calculated for all the windows.
Short Term Feature Extraction : Extract Mel Frequency Cepstral Coefficients (MFCCs). A 128-mixture GMM is then trained to model the coefficients, with MAP adaptation applied to update the means and weights for all mixtures. The means are then concatenated to form a 1536 (12*128 mixtures) dimension supervector (SPV)
Dimensionality Reduction : Apply dimensionality reduction algorithm on the 1536 dimensional short-term feature supervector to 800 dimensional supervector.
Offline Training (step 1) : Train SVM1, SVM2, SVM4, SVM5 and the age regressors with 70% of the data.
Offline Training (step 2) : Use 15% of the data two generate outputs of the classifiers and the regressors trained in step 1 and use this to train SVM3.
Testing : Age Group and Gender of a speaker is predicted .

We further see that use of WSNMF described in section 3 to compress the 1536 features used in short term feature supervector to 800 features leads to increase in the prediction accuracy. With the compressed features the accuracy of classifier based on short term feature SPV increased from 69.7% to 75.8% and the overall accuracy by combining outputs of all 3 classifiers was improved from 73.5% to 79.5%.

A. Accuracy Measure

We have used the standard accuracy metric

$$Accuracy = \frac{1}{M}\sum_{i=1}^{M}\mathbb{I}[\hat{y} = y] * 100 \quad (8)$$

where, $\hat{y}$ denotes the predicted age group and y denotes the actual age group.

TABLE II. Accuracy

Differnt Classifiers	Accuracy
SVM1	65.0
SVM2	69.7
SVM1 and SVM2 combined	70.45
SVM1, SVM2, SVM3 combined	73.5
SVM2 and WSNMF	75.8
combined and WSNMF	79.5

B. Confusion Matrices

In a confusion matrix [**?**] we list the true class along the rows and the prdicted class along the colunm. The (i,j) entry of the matrix denotes the number of samples belonging to class "i" that have been classifed as class "j". The confusion matrices for 6-class classifiers and gender classifiers are

Gender Class	M	F
M	63	3
F	1	65

(Using Long Term Features)

Gender Class	M	F
M	61	5
F	12	55

(Using Short Term Features)

Age Class	1	2	3	4	5	6
1	19	2	0	1	0	0
2	7	10	1	2	3	0
3	8	3	8	0	2	0
4	0	0	0	19	1	0
5	0	0	0	5	9	6
6	0	1	0	3	0	22

(Using Long Term Features)

Age Class	1	2	3	4	5	6
1	19	0	2	0	1	0
2	3	19	1	0	0	0
3	2	6	12	0	0	1
4	0	0	0	15	1	4
5	0	4	1	4	7	4
6	0	0	4	0	2	20

(Using Short Term Features)

Age Class	1	2	3	4	5	6
1	19	0	3	0	0	0
2	5	13	4	0	0	1
3	0	4	17	0	0	0
4	0	0	0	20	0	0
5	0	0	0	5	5	10
6	0	0	1	1	1	23

(Combined Classification using all 3 classifiers)

Fig. 2. Figure illustrating Confusion Matrices for gender and age group detection using Short Term Features , Long Term Features and both.

VII. Conclusion

In this paper, we use the pitch and formant information in the form of Long term features(Mean, standard deviation etc.) and the 12 Mel-Frequency Cepstral coefficients in the form of means of 128 mixture Gaussian mixture model(STF-SPV). These features are used to train a classification model. We have shown that accuracies of about 73.5% is achievable under this framework for 6 class classification.

We further demonstrate that use of compression tequniques on the 1536 dimensional short term feature supervector(STF-SPV) can lead to increase in the prediction accuracy. We use WSNMF to reduce the dimensionality of STF-SPV from 1536 to 800 and this results in increase in prediction accuracy from 73.5% to 79.5%.

VIII. Fututre work

The limitation of this project is unavailibilty of large and properly labelled training data. The results of this project are hence a good indicator towards the possibility of achieving the goal of age-gender classification with high accuracy. However it needs to be verified on larger datasets to get the exact estimate of achievable accuracy.

Further, the results of the project indicate that use of compression tecniques specifically WSNMF can lead to increase in the prediction accuracy and hence this needs to be explored further.

References

[1] E. Ycesoy and V. V. Nabiyev, "Age and gender recognition of a speaker from short-duration phone conversations," in *Signal Processing and Communications Applications Conference (SIU), 2015 23th*, May 2015, pp. 751–754.

[2] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, March 2008, pp. 1605–1608.

[3] E. Fazl-Ersi, M. E. Mousa-Pasandi, R. Laganire, and M. Awad, "Age and gender recognition using informative features of various types," in *Image Processing (ICIP), 2014 IEEE International Conference on*, Oct 2014, pp. 5891–5895.

[4] Open speech data corpus for german. [Online]. Available: https://www.researchgate.net/publication/264858885$_{I}dentification_{o}f_{A}ge-Group_{f}rom_{C}hildren's_{S}peech_{b}y_{C}omputers_{a}nd_{H}umans$

[5] D. Mahmoodi, H. Marvi, M. Taghizadeh, A. Soleimani, F. Razzazi, and M. Mahmoodi, "Age estimation based on speech features and support vector machine," in *Computer Science and Electronic Engineering Conference (CEEC), 2011 3rd*, July 2011, pp. 60–64.

[6] C. van Heerden, E. Barnard, M. Davel, C. van der Walt, E. van Dyk, M. Feld, and C. Mller, "Combining regression and classification methods for improving automatic speaker age recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 5174–5177.

[7] M. H. Bahari and H. V. hamme, "Speaker age estimation using hidden markov model weight supervectors," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, July 2012, pp. 517–521.

[8] Open speech data corpus for german. [Online]. Available: http://www.voxforge.org/home/forums/other-languages/german/open-speech-data-corpus-for-german