Improved Speaker Age Group and Gender Detection using Multiple Classifiers

Abhishek K. Singh, Abhay Kumar, Tanushree Gupta, Ayushi Singhal
Under the guidance of Dr. R. M. Hegde and Ajay Dagar

Motivation

- Paralinguistic cues such as identity, gender, perceptual age, etc. can be perceived from the utterances of people.

- Such info. can be used in computer recommendation systems and in human-computer interaction.

Problem Description

- Given voice samples of speakers, efficiently classify them on the basis of their age and gender

- Sub-Problems -

1. Extract appropriate features which can be used to distinguish between age-groups and genders
2. Make suitable system of multiple classifiers for efficient classification

Outline of the Approach

- Long Term Features (LTF) and Short Term Features (STF) have been extracted from voice samples.
- STFs have also been dimensionally reduced using Weighted Supervised Non-negative Matrix Factorization (WSNMF).
- Feature vectors have been fed to SVMs and Regressors to get the final age and gender estimate of the speaker.

Feature Extraction

Long Term Features

- Pitch (using ACF) and the first three formants (using Burg's Algorithm) were extracted.
- The LTF feature vector then consists of the mean, max., min., difference, and std. deviation across estimates of the pitch and the 3 formants.

Feature Extraction

Short Term Features

- 12 MFCCs were extracted for all the utterances.
- Each MFCC set was 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 were then combined to form a 12 coefficients X 128 mixture i.e. 1536 dimensional supervector.

Weighted Supervised Non-negative Matrix Factorization

- NMF is used to factorize a non-negative matrix V into two lower dimensional non-negative matrices W and H.
- These matrices can be found by minimizing the Kullbeck-Liebler distance.
- The STF supervector (dim. = 1536) can be dimensionally reduced using WSNMF to find H (dim. = 800) and hence dim. of the input space for training can be reduced.

Corpus and Classification

Open Speech Data Corpus of the German speaker is used for training and testing of the system. It consists of 5 voice sample of each 874 german speaker. These speakers are evenly distributed among all the age classes and gender.

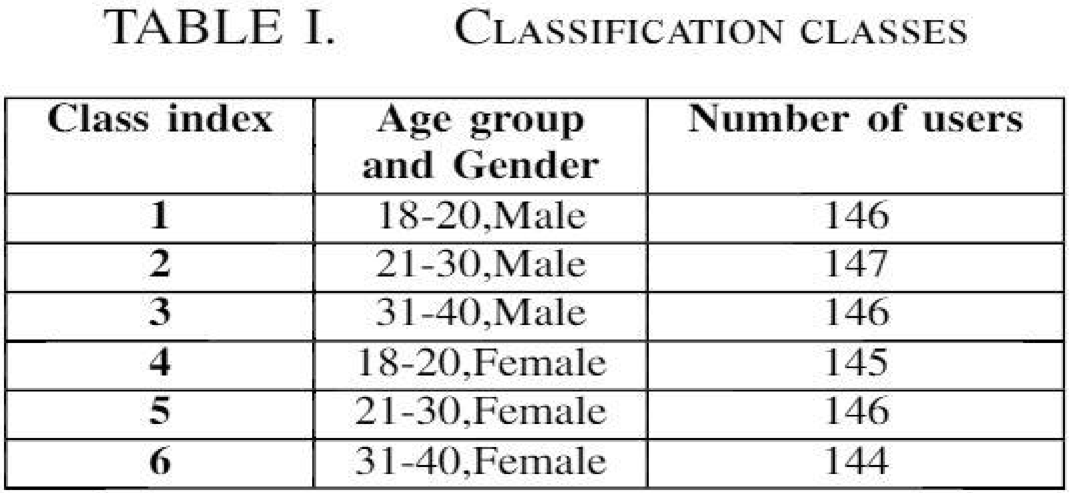
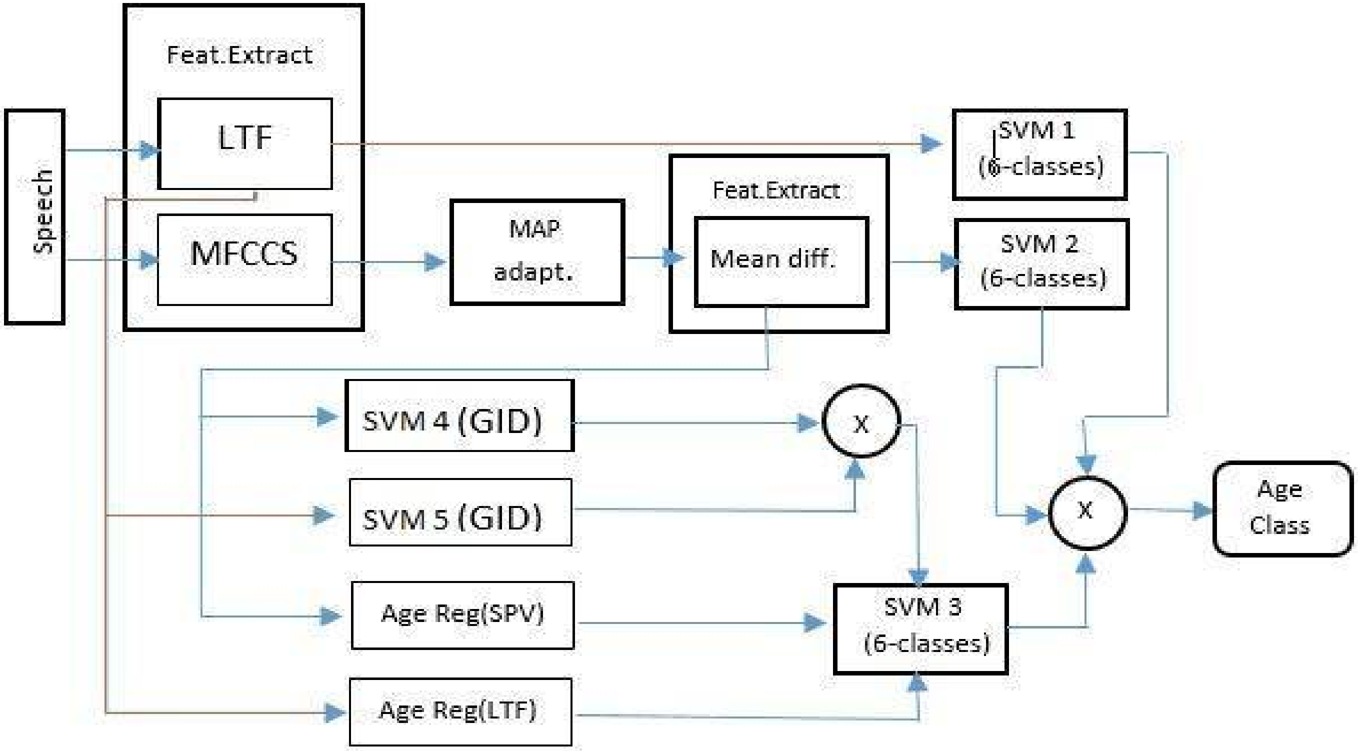


TABLE I. CLASSIFICATION CLASSES

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

Classification Design



Performance Evaluation

Different Classifiers	Accuracy(%)
SVM1	65.0
SVM2	69.7
SVM1 and SVM2 combined	70.45
SVM1, SVM2 and SVM3 combined	73.5
SVM2 and WSNMF	75.8
SVM1,SVM2,SVM3 and WSNMF	79.5

Conclusion

- 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) has been used to train a classification model.
- An accuracy of about 73.5% is achievable under this framework for 6 class classification.
- Compression technique (WSNMF) on the 1536 dimensional short term feature supervector(STF-SPV) to dim. 800 vector leads to increase in the prediction accuracy from 73.5% to 79.5% .

Future work

- A limitation of this project is unavailabilty of large and properly labelled training data. thus the results need 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 techniques 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 super vectors 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, Oct2014, pp. 5891–5895.

[4] 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.

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

Thank you

Improved Speaker Age Group and Gender Detection using Multiple Classifiers

Abhishek K. Singh, Abhay Kumar, Tanushree Gupta, Ayushi Singhal

Under the guidance of Dr. R. M. Hegde and Ajay Dagar

Motivation

- Paralinguistic cues such as identity, gender, perceptual age, etc. can be perceived from the utterances of people.

- Such info. can be used in computer recommendation systems and in human-computer interaction.

Problem Description

- Given voice samples of speakers, efficiently classify them on the basis of their age and gender
- Sub-Problems -

1. Extract appropriate features which can be used to distinguish between age-groups and genders
2. Make suitable system of multiple classifiers for efficient classification

Outline of the Approach

- Long Term Features (LTF) and Short Term Features (STF) have been extracted from voice samples.
- STFs have also been dimensionally reduced using Weighted Supervised Non-negative Matrix Factorization (WSNMF).
- Feature vectors have been fed to SVMs and Regressors to get the final age and gender estimate of the speaker.

Feature Extraction

Long Term Features

- Pitch (using ACF) and the first three formants (using Burg's Algorithm) were extracted.
- The LTF feature vector then consists of the mean, max., min., difference, and std. deviation across estimates of the pitch and the 3 formants.

Feature Extraction

Short Term Features

- 12 MFCCs were extracted for all the utterances.
- Each MFCC set was 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 were then combined to form a 12 coefficients X 128 mixture i.e. 1536 dimensional supervector.

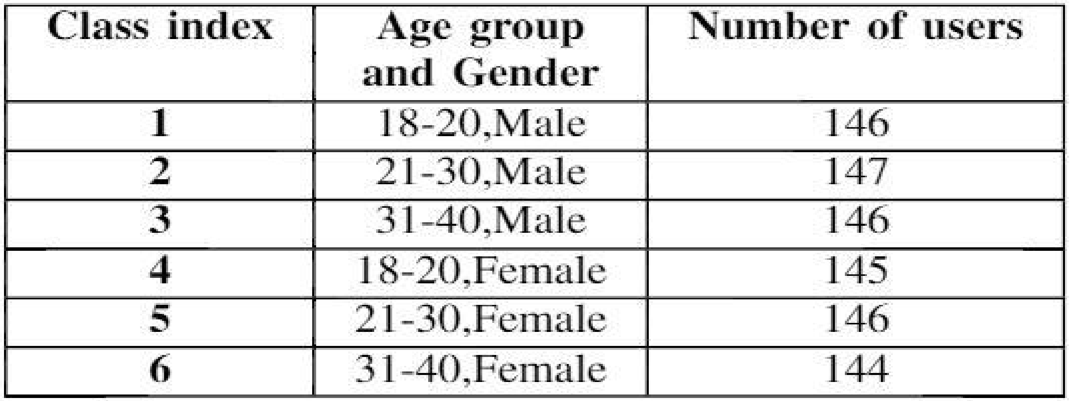
Weighted Supervised Non-negative Matrix Factorization

- NMF is used to factorize a non-negative matrix V into two lower dimensional non-negative matrices W and H.
- These matrices can be found by minimizing the Kullbeck-Liebler distance.
- The STF supervector (dim. = 1536) can be dimensionally reduced using WSNMF to find H (dim. = 800) and hence dim. of the input space for training can be reduced.

Corpus and Classification

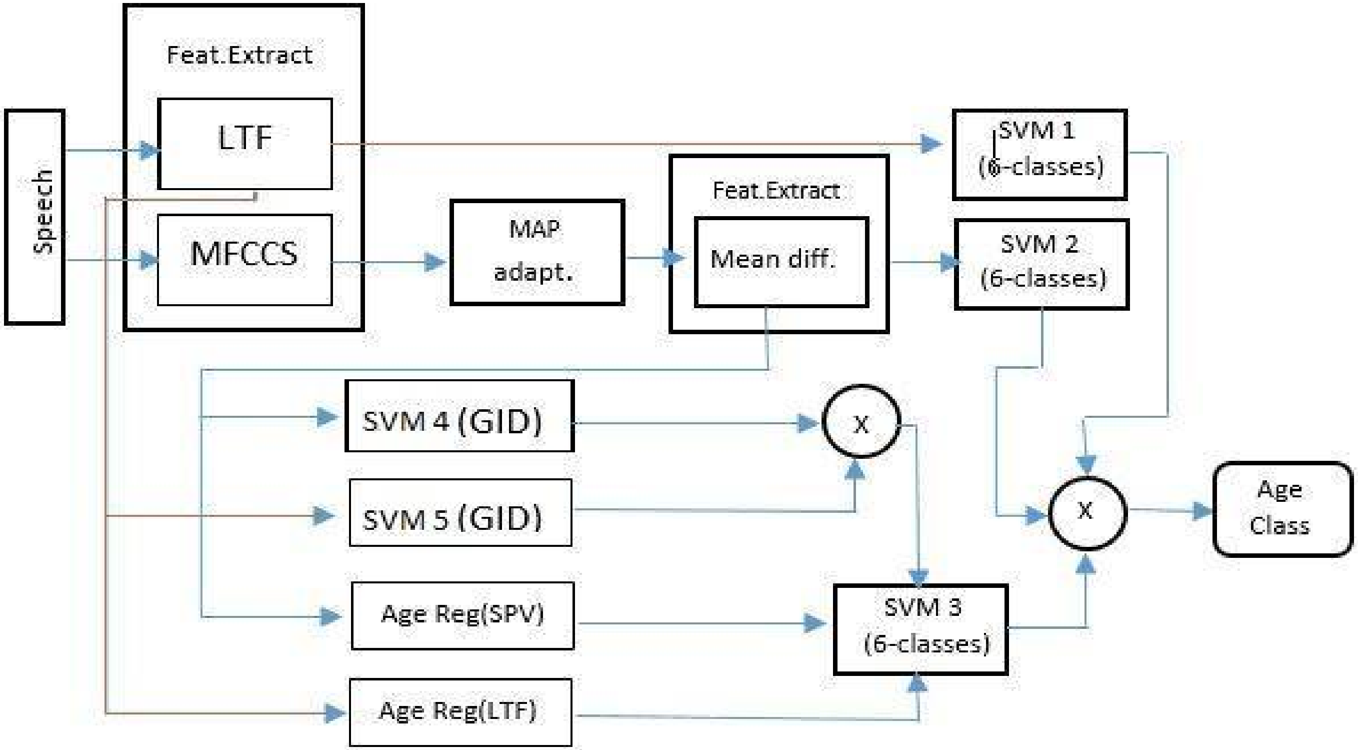
Open Speech Data Corpus of the German speaker is used for training and testing of the system. It consists of 5 voice sample of each 874 german speaker. These speakers are evenly distributed among all the age classes and gender.

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

Classification Design



Performance Evaluation

Different Classifiers	Accuracy(%)
SVM1	65.0
SVM2	69.7
SVM1 and SVM2 combined	70.45
SVM1, SVM2 and SVM3 combined	73.5
SVM2 and WSNMF	75.8
SVM1,SVM2,SVM3 and WSNMF	79.5

Conclusion

- 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) has been used to train a classification model.
- An accuracy of about 73.5% is achievable under this framework for 6 class classification.
- Compression technique (WSNMF) on the 1536 dimensional short term feature supervector(STF-SPV) to dim. 800 vector leads to increase in the prediction accuracy from 73.5% to 79.5% .

Future work

- A limitation of this project is unavailabilty of large and properly labelled training data. thus the results need 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 techniques 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 super vectors 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, Oct2014, pp. 5891–5895.

[4] 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.

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

Thank you