Speaker Recognition with Emotional Speech

Hussain Ahmad Faraz, He Qianhua, Rehman Ullah
South China University of Technology

Abstract

In recent researches, emotional speaker recognition has emanated as an important challenging topic. Despite the fact that speaker recognition research has been ongoing for extra than four decades, the speaker recognition performance is effected by background noise, age, person health and emotional state of a speaker. I-vector is used in this study because it has been proved to be very efficient for its fixed length and low dimensions. For channel/session compensation, Linear Discriminant Analysis (LDA) and Probabilistic Linear Discriminant Analysis (PLDA) are used. In the experiments, text-dependent CREMA-D database and text-independent database are used. Satisfying results are achieved in noisy environment.

Methodology

The framework of proposed emotional speaker recognition system is instantiated in figure 1. CREMA-D database that consist of six emotions is used in both clean and noisy environments. Real subway noise is added to emotional speech utterances to obtain noisy CREMA-D database. MFCC is used to extract the most relevant information from the emotional speech signals to represent them in feature vectors. Gaussian Mixture Models (GMM) is used as a tool for i-vector. I-vector is very popular in speaker recognition field. A low dimensional of 400 dimensions is used. For channel compensation, LDA and PLDA techniques are used.

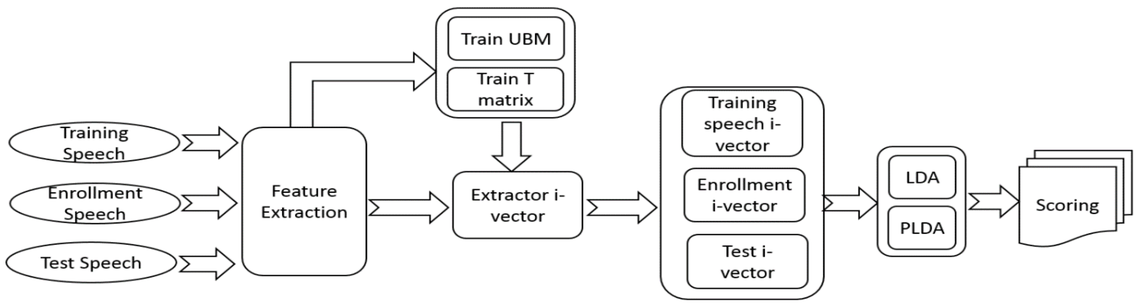

Fig. 1. Emotional speaker recognition system

Results

Kaldi toolkits D. Povey, 2011 used for performing experiments. MFCC features extraction (20 ms hamming window, every 10ms), 19 Mel-frequency cepstral coefficient together with log energy were used. Delta and delta-delta coefficient were evaluated to generate 60-dimensional feature vector.256 Gaussian Mixtures, 400-dimensional i-vector and 150-dimensional LDA/PLDA.GPU: GTX 1080T used for our experiments. Text-dependent and text-independent emotional speaker recognition results are shown in Table 1 and 2.

Table 1: The EERs (%) of Emotional Speaker Verification using GMM: 256 & PLDA

EMOTION	ANG	DIS	FEA	HAP	NEU	SAD
EER (%)	2.857	2.418	3.58	2.198	1.758	3.297

Table 2: The EERs (%) of Emotional Speaker Verification using GMM: 256 & PLDA

EMOTION	ANG	FEA	HAP	NEU	SAD
EER (%)	11.77	16.19	13.14	8.209	17.43

Conclusion

In this study an emotional speaker verification system presented in which the factor analysis is done by low-dimensional space which consists of both speaker and channel variabilities.. To compensate the intersession problem, two different techniques like LDA and PLDA were used. CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) emotional database is used for our experiments. The performance of PLDA is better than LDA using GMM: 256. Experimental results show that our model performed well under clear. Neutral emotion showed best result. After performing these experiments, our model was checked on noisy database. The results were not as good as in clear environment but still reasonable. Neutral emotion with 5 EER has the best result.

Acknowledgements

Funding: This researched has been completed without any funding.
Conflict of Interest: The authors declare that they have no conflict of interest.