Speaker Identification in a MultiSpeaker Environment

Manthan Thakker Prachi Ved Shivangi Vyas Sahil Kathpal
Thadomal Shahani Engineering College.

Contact Information:
IT Department,
Thadomal Shahani Engineering College
Bandra(West),Mumbai

India
Phone: +91-9892059543
Email: manthanthakker40@gmail.com

Abstract

The human auditory system is capable of performing many interesting tasks, several of which could find useful applications in engineering settings. One such capability is the ability to perceptually separate sound sources, allowing a listener to focus on a single speaker in a noisy environment. This effect is often referred to as the cocktail effect (in reference to a cocktailparty environment where several simultaneous conversations are taking place in the background). This paper introduces methodologies for identifying a desired speakers audio stream from a binaural recording of multiple speakers in conversation. Our proposed work consists of truncating a recorded voice signal, framing it, passing it through a window function, calculating the Short Term FFT, extracting its features and matching it with a stored template. Cepstral Coefficient Calculation and Mel frequency Cepstral Coefficients(MFCC) are applied for feature extraction purpose. Gaussian Mixture Model (GMM) is constructed for analysing overlapped speech segments.

Introduction

Speaker recognition is the process of automatically recognizing who is speaking by using the speaker-specific information included in speech waves to verify identities being claimed by people accessing systems; that is, it enables access control of various services by voice.Our project does an Gaussian Mixture Component Analysis on the Cocktail Party Problem i.e. there are several speakers, and a microphone in a room. Overlapping speech, which means multiple persons speak simultaneously, occurs frequently in natural conversation. The goal is to identify the voices of individual speakers from the different voices from the signals recorded from the microphone.Figure 1.1 shows our basic strategy for speaker identification.

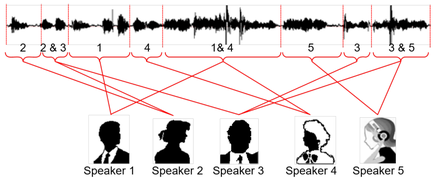

Figure 1: Illustration of our speaker-identification task.

Methods and Procedures

The paper uses Mel-scale Frequency Cepstrum Coefficients (MFCC) as a base for matching speaker samples to the test signal.The proposed system comprises of two states, a training state, and a test/verification state: The figure below charts this process:

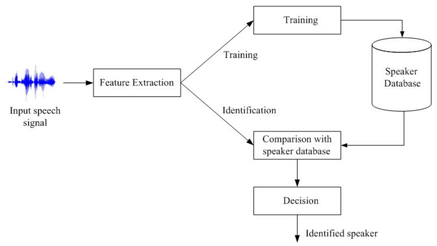

Figure 2: Testing and verification stages.

The first step of training is to transform the speech signal to a set of vectors, and to obtain a new acoustic parametric representation of the speech which is more suitable for statistical modeling. Firstly the speech signal is broken up into short frames of 25 to 30 ms, then windowed to minimize distortion. The signal is then analyzed and stored as that users template.The first step in the testing stage is to extract features from the input speech, similar to that during training, compare the input speech to all other stored templates and select the most accurately matching template and ID the speaker.

Phase 1

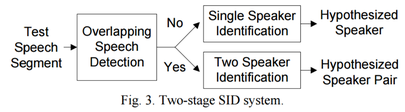

Fig. 3. Two-stage SID system.

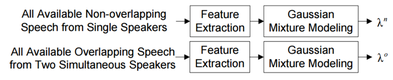

Figure 3: Identifying single or multi-speaker.

A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation,

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i \; g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), wi , i = 1, . . . , M, are the mixture weights, and g(x—i, i), i = 1, . . . , M, are the component Gaussian densities. Each component density is a D-variate Gaussian function.

Results of Phase 1

On performing the MFCC analysis for 4 speakers and calculating the mean and standard deviations of the speech signals, the following result was obtained:

	Speaker 1 :												
mean	-0.7628	10.8294	2.1436	4.3120	6.6074	1.6028	-1.6155	-0.0644	-0.7000	-2.2491	-0.8980	-0.0780	-0.0417
std dev	10.1117	8.2131	5.3377	4.9654	6.6308	5.2013	4.8729	3.0562	2.5929	2.6398	1.6964	1.0151	0.3873
	Speaker 2:												
mean	-0.7881	10.0581	5.5021	5.0647	2.1692	1.3871	0.4515	-0.9620	-0.9008	-0.8938	-0.5831	-0.3374	-0.0860
std dev	10.1007	9.2671	6.5762	5.9658	6.1702	3.6106	3.2361	2.9000	2.3220	1.8247	1.3538	0.7017	0.2559
	Speaker 3:												
mean	-1.2476	9.7126	4.0755	4.9790	3.0286	1.1762	0.6040	0.1841	-0.3504	-0.5027	0.0836	-0.1024	-0.0082
std dev	9.8278	9.2259	6.2314	6.1002	6.5868	4.7615	4.1078	2.9686	2.7108	1.8778	1.3315	0.8623	0.2907
	Speaker 4:												
mean	0.2585	11.4502	5.7913	5.4526	1.7112	0.8895	-0.4191	-0.8870	-0.3804	-0.5787	-0.1351	-0.2491	-0.1111
std dev	10.9549	10.8168	7.1071	6.7188	7.1513	4.3714	3.7725	3.4143	2.4483	2.0386	1.5109	0.6340	0.2421

Figure 4: MFCC vectors

The following vectors will further help us in deriving classification and clustering values for speaker segmentation.

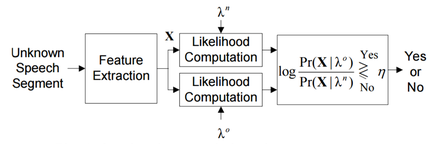

Fig. 4. Overlapping speech detection.

Phase 2

The following concept will provide the basis for the identification of individual speakers in overlapping speech segments.

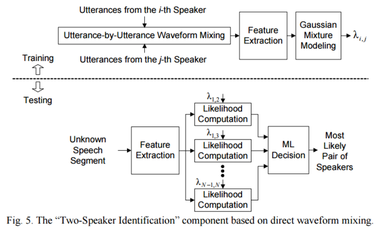

Fig. 5. The "Two-Speaker Identification" component based on direct waveform mixing.

Conclusions

- Most of the speaker-identification systems proposed until now only focus on the identification of single speaker; however, in reality, many audio recordings involve multiple persons speaking simultaneously. This paper examined the feasibility of detecting and identifying speakers with overlapping parts in speech. Our encouraging results arrived at this initial stage of investigation laid a good foundation for the future development of robust speaker-identification system that works for conversation or meeting recordings consisting of multiple overlapping speakers voices. To be of more practical use, it is needed to scale up the system to handle a wider variety of speech data and speaker population size in the future. In addition, we will investigate the problem of speaker diarization for overlapping speech, based on unsupervised speaker segmentation and clustering.

Forthcoming Research

The proposed system would serve as the basis for speaker speech isolation. One can also try to increase the accuracy of the results by using multiple feature extraction techniques such as wavelet transformations.

Acknowledgements

Prof. Shanthi Therese, Our Project Guide.

References

1. D. Reynolds and R. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Transactions on Speech Audio Processing,Vol. 3, 1995, pp. 72-83.
2. Wei-Ho Tsai and Shih-Jie Liao, Speaker Identification in Overlapping Speech, Journal Of Information Science and Engineering paper published in 2010.