THE INTERNATIONAL JOURNAL OF
SCIENCE & TECHNOLEDGE

Comparative Study of Speaker Segmentation Using MFCC and LSF

Kiran C. Jamgade
Student, Department of Electronics and Telecommunication, Babasaheb Naik College of Engineering, Maharashtra, India
Dr. Naresh P. Jawarkar
Professor &HOD, Department Electronics and Telecommunication, Babasaheb Naik College of Engineering, India

Abstract:
Speaker segmentation is defined as the process by which a speech signal of long duration is partitioned into homogenous segments by detecting changes of speaker identity. Speaker segmentation algorithms can be broadly classified into three categories: model based, metric based and hybrid. In the model-based segmentation, a set of models is derived and trained for different speaker classes from a training corpus. The incoming speech streams are classified using these models [4]. However, in many cases, the pre-knowledge of speakers and acoustic classes are often not available. Some of the model based classification methods are Gaussian Mixture Models (GMM), GMM with multilayer perceptron (MLP), K-nearest neighbour (KNN), Support Vector Machines (SVM), etc. [5].The objective of the paper is to implement speaker diarization system using distance metric. The database used are the recorded files that contains random number of male and female voice and nonspeech signals such as music, noise etc. The feature extraction process include two techniques: Mel Frequency Cepstral Coefficients(MFCC) and Linear Spectral Frequencies(LSF). For segmentation Hotelling T2-statistic and Bayesian Information Criterion (BIC) technique are used. In segmentation coarse segmentation is carried out with T2 distance and for refinement and confirmation of speaker change BIC distance is used. In this paper, a speaker segmentation system has been performed on recorded data using feature extraction methods such as mel frequency cepstral coefficient (MFCC), linear spectral frequency (LSF) and models for Distance Calculations like T2 and Bayesian Information Criterion (BIC), and thereafter classifying a signal into segments. Finally, we present an analysis of speaker segmentation performance as reported through the MFCC and LSF on and identify important areas for future research. And precision, recall and figure of merit are calculated.

1. Introduction

Speaker segmentation is defined as the process by which a speech signal of long duration is partitioned into homogenous segments by detecting changes of speaker identity. Speaker segmentation algorithms can be broadly classified into three categories: model based, metric based and hybrid. In the model-based segmentation, a set of models is derived and trained for different speaker classes from a training corpus. Some of the model based classification methods are Gaussian Mixture Models (GMM), GMM with multilayer perceptron (MLP), K-nearest neighbour (KNN), Support Vector Machines (SVM), etc. [5]. Metric-based method accesses the similarity between neighbouring analysis windows over the audio stream by a distance function of their metric. A wide variety of distance metrics could be used. A commonly used metrics are Bayesian Information Criterion (BIC), second order Hotelling'sT^2 statistics [7] and the Kullback-Leibler divergence (Gaussian divergence) [8]. The hybrid method uses both model-based and metric-based approach. Various techniques used for speech, non-speech detection are Line Spectrum and pitch features to detect change points, Perceptual linear prediction (PLP) cepstral coefficients, average zero crossing rate [8]. Speaker change can be detected using different features. Lu and Zhang [9] have used a multifeature set consisting of Mel frequency cepstral coefficient (MFCC), Line Spectrum and pitch features to detect change points. Perceptual linear prediction (PLP) cepstral coefficients are used in [10]. R. Huang et al. [10] have considered perceptual based minimum variance distortionless response (PMVDR), smooth zero crossing rate (SZCR), filter bank log energy coefficients (FBLC). Jawarkar et al. have used line spectral pair (LSP) features for speaker segmentation For speaker change detection Hotelling T2 distance and Baysian information criteria is used. Here in this paper zero crossing rate is used to detect speech non speech signal. The objective of this paper is developing the database for speaker segmentation, extracting the features using MFCC and LSF and then developing speaker segmentation. Next, section deals with system architecture, in which speech detection is carried out with average zero crossing rate and then feature extraction process is implementated. Then it represents the experimental work. The database used, observations and results are discussed in this paper. Lastly it represents conclusions and the future scope.

2. System Architecture

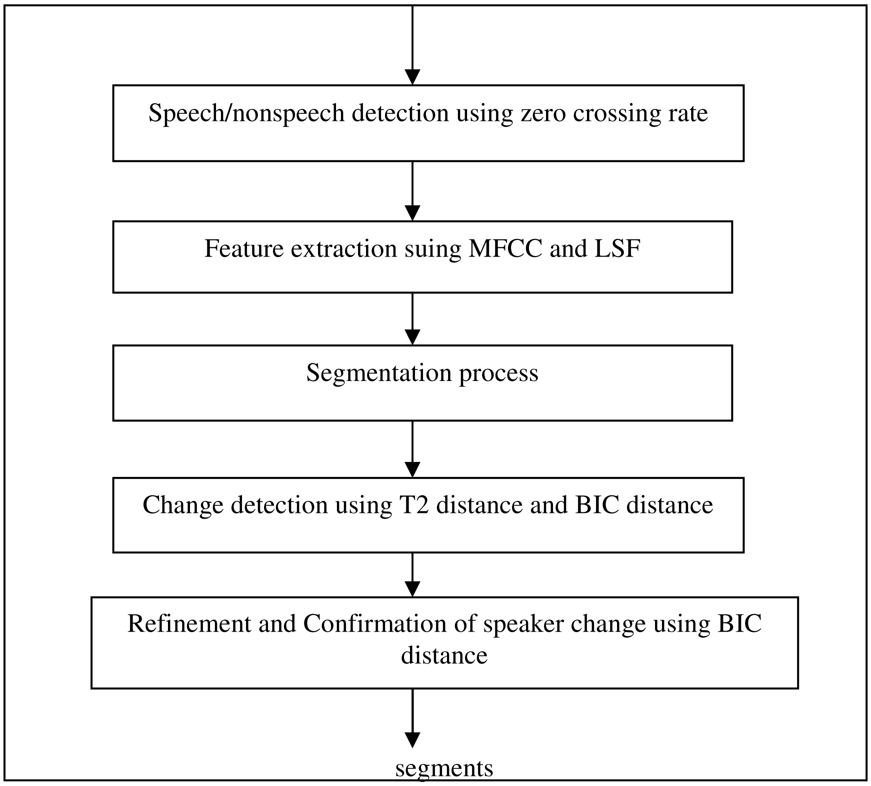
Speech/nonspeech detection using zero crossing rate
Feature extraction suing MFCC and LSF
Segmentation process
Change detection using T2 distance and BIC distance
Refinement and Confirmation of speaker change using BIC distance
segments

Figure 1: system architecture

The architecture of the system is as shown in figure1. The basic blocks of speaker segmentation system are speech signal, speech detection, change detection, segmentation.

2.1. Speech Detection

The aim of this step is to find the regions of speech in the audio stream. Depending on the domain data being used, nonspeech regions to be discarded can consist of many acoustic phenomena such as silence, music, room noise, background noise or cross-talk.[3]

The general approach used is maximum likelihood classification with Gaussian Mixture Models (GMMs) trained on labeled training data, although different class models can be used, such as multi-state HMMs. The simplest system uses just speech/non-speech models. Various techniques used for speech, non-speech detection are Line Spectrum and pitch features to detect change points, Perceptual linear prediction (PLP) cepstral coefficients, average zero crossing rate[8].

In this paper average zero crossing rate is used for f speech/nonspeech detection:

2.2.1. Average Zero Crossing Rate

ZCR is the number of zero-crossings within the frame The signal is sampled at the sampling frequency of 22050 Hz. The silence removal stage removes the silence portion of the signal based on the energy threshold criterion. ZCR is calculated as under :

$$z(m) = \frac{1}{N}\sum_{n-m-N+1}^{m} \frac{|sgn(s(n))-sgn(s(n-1))|}{2} \quad w(m-n) \qquad (2.1)$$

where N is the length of the frame, m is the endpoint of the frame, and $w(n)$ is the window function Therefore, each 2 s-audio results in a speech/nonspeech decision task as follows:

$$d_i = \begin{cases} z_i & z_i \geq \theta \\ 0 & otherwise \end{cases} \qquad (2.2)$$

where 1 means speech, 0 means nonspeech, z_i is the average ZCR value in the ith audio frame and θ is the threshold.

2.2. Feature Extraction

The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier. Different approaches and various kinds of audio features were proposed with varying success rates. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. Several feature extraction algorithms can be used to do this task, such as - Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCC), and Human Factor Cepstral Coefficient (HFCC).

The MFCC algorithm is used to extract the features. The functions used for feature extraction. MFCC are chosen for the following reasons: -

- MFCC are the most important features, which are required among various kinds of speech applications.
- It gives high accuracy results for clean speech.
- MFCC can be regarded as the "standard" features in speaker as well as speech recognition.

2.2.1. MFCC (Mel Frequency Cepstral Coefficients)

Only voiced segments of speech signal are processed for MFCC extraction. The procedure to determine MFCC is described as follows:

i. Segmenting all concatenated voiced speech signal into 25.6ms-length frames.
ii. Estimating the logarithm of the magnitude of the discrete Fourier Transform (DFT) for all signal frames.
iii. Filtering out the center frequencies of the sixteen triangle band-pass filters corresponding to the mel frequency scale of individual segments.
iv. Estimating inversely the IDFT to get all MFCC coefficients.

The melscale used in this work is to map between linear frequency scale of speech signal to logarithmic scale for frequencies higher than 1 kHz. This makes the spectral frequency characteristics of signal closely corresponding to the human auditory perception.

- Pre-emphasis: The continuous time signal (speech) is sampled at sampling frequency. At the first stage in MFCC feature extraction is to boost the amount of energy in the high frequencies. This pre-emphasis is done by using a filter.

$$y(n) = x(n) - \alpha x(n-1) \qquad (2.3)$$

where *x(n)* is the input speech signal. The recommended values of pre-emphasis factor α are in the range [0.95, 0.98]. In the present study, α was chosen as 0.97.

- Windowing: Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame.

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{2N-1}\right), & 0 \le n \le N-1 \\ 0, & \text{otherwise} \end{cases} \qquad (2.4)$$

- Mel Filter bank and Frequency wrapping: The mel filter bank consists of overlapping triangular filters with the cut-off frequencies determined by the centre frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale.

$$f_{mel} = 2595\log_{10}\left(1 + \frac{f}{700}\right) \qquad (2.5)$$

where f_{mel} is Mel frequency and f is the frequency in Hz. This led to the definition of MFCC.

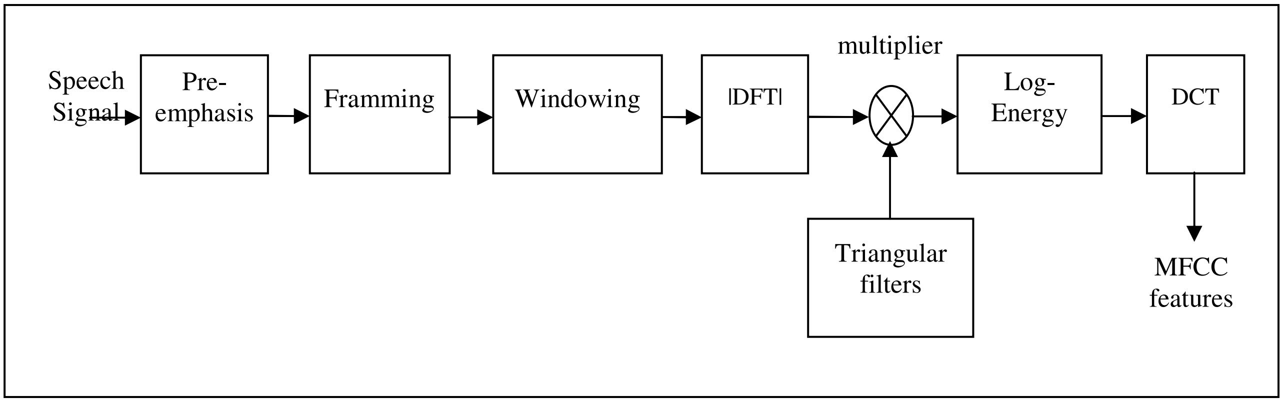

Figure 2: MFCC calculation

- Take Logarithm: The logarithm has the effect of changing multiplication into addition. Therefore, this step simply converts the multiplication of the magnitude in the Fourier transform into addition.

- Take Discrete Cosine Transform: It is used to orthogonalise the filter energy vectors. Because of this orthogonalization step, the information of the filter energy vector is compacted into the first number of components and shortens the vector to number of components.

2.2.2. LSF (Line Spectral Frequency)

Line spectrum frequency, LSF, is a way of uniquely representing the LPC-coefficients. The motivation behind LSF transformation is greater interpolation properties and robustness to quantization. These benefits are achieved by the cost of higher complexity of the overall system. It has many interesting properties such as (i) all zeros of LSF polynomials are on the unit circle, (ii) the corresponding zeros of the symmetric and ant symmetric LSF polynomials are interlaced, and (iii) the reconstructed LPC all-pole filter preserves its minimum phase property if (i) and (ii) are kept intact through a quantization procedure.

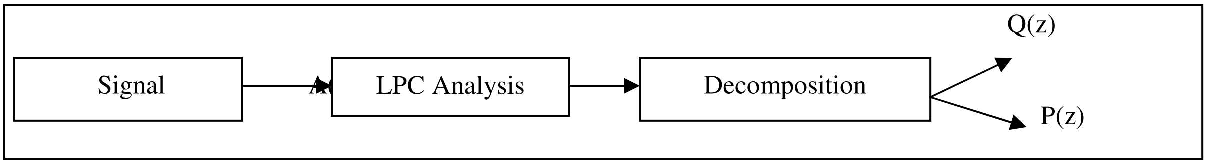

Figure 3: Decomposition of A(z).

The transfer function of a simplified vocal tract model based on P^{th} order LPC speech analysis is given by

$$H(z) = \frac{1}{1 + \sum_{p=1}^{P} a_p z^{-p}} = \frac{1}{A(z)} \qquad (2.9)$$

Polynomials $P(z)$ &$Q(z)$ are formed as follows:

$$P(z) = A(z) + z^{-(P+1)} A(z^{-1}) \qquad (2.10)$$

$$Q(z) = A(z) - z^{-(P+1)} A(z^{-1}) \qquad (2.11)$$

$P(z)$is a symmetric polynomial while $Q(z)$is an anti-symmetric polynomial and

$$A(z) = \frac{P(z) + Q(z)}{2} \qquad (2.12)$$

Steps involved in LSF computation are:

- The speech signal is divided into frames and LPC coefficients are determined for each frame to form $A(z)$**.**
- The roots of polynomials $P(z)$ and $Q(z)$ are determined.
- The phases $\theta_p(i)$ and $\theta_q(i)$ from roots of $P(z)$ and $Q(z)$, respectively, are determined.
- Finally, the line spectrum frequencies are determined using the following equations.

$$f_p(i) = \frac{\theta_p(i)}{2\pi} \qquad (2.13)$$

$$f_q(i) = \frac{\theta_q(i)}{2\pi} \qquad (2.14)$$

In the present study 10th order LPC were computed using autocorrelation method for each frame to determine LSF.

3. Speaker Change Detection

The aim of this step is to find points in the audio stream likely to be change points between audio sources. If the input to this stage is the un-segmented audio stream, then the change detection looks for both speaker and speech/nonspeech change points [7][8]. There are many distance matrices that can be used to detect speaker change and they are Hotelling T2-statistic**,** Bayesian Information Criterion (BIC) technique KL-2 distance. In this project for the speaker change detection, two techniques are used that are Hotelling T2-statistic **and** Bayesian Information Criterion (BIC) technique [8]. In this paper, Hotelling T2- statistic was used for coarse segmentation and Bayesian Information Criterion (BIC) was used for refinement and speaker change confirmation.

➢ Hotelling T2-statistic: To detect the potential change over point from the local peaks in the dissimilarity sequence, they must satisfy the following conditions:

(i) T2(i) > T2(i-1)
(ii) T2(i) > T2(i+1)
(iii) T2(i) > Threshold

Hotelling T2-statistic for speaker segmentation is that for two audio segments, if they can be modeled by multivariate Gaussian distributions: N (μ_1, Σ_1) and, N (μ_2, Σ_2) we assume their covariance are equal but unknown.

$$T2 = (ab/a+b)\,(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) \qquad (2.15)$$

Where a and b are the number of frames within each of the audio segments, respectively. And Σ is the covariance matrix for window of size (a+b) frames. If the processing audio window is shorter than 2 s, even a global covariance will suffer from insufficient estimation.

- Bayesian Information Criterion (BIC) technique:

The first general approach used for change detection, used is a variation on the Bayesian Information Criterion (BIC) technique [8]. This technique searches for change points within a window using a penalised likelihood ratio test of whether the data in the window is better modeled by a single distribution (no change point) or two different distributions (change point).

Let N (μi, Σi) and N(μj ,Σj) be the Gaussian models derived from two speech segments and Ni and Nj be the corresponding number of feature vectors used in the estimation. Let also N(μ,Σ) be a single Gaussian model estimated on the union of the aforementioned speech segments, the BIC difference between the two models can be defined as:

$$\Delta BIC(\Sigma_i,\Sigma_j) = \frac{1}{2}((N_i + N_j) \log |\Sigma| - Ni \log |\Sigma_i| - Nj \log |\Sigma_j|) - \frac{1}{2} \lambda(\delta + \frac{1}{2}\delta(\delta+1)) \log(N_i + N_j) \qquad (2.16)$$

N_i, N_j are no. of feature vector in 1st and 2nd Gaussian model, respectively.Σ_i, $\Sigma_j$arecovariance matrix of respective Gaussian model. λ is thePenality factor , δis the feature vector dimension.

4. Experimental Work

Database used consists of five files: file1(1M-1F), file2(3F), file3(3M2F), file4(FFF). File5(FFF). Each file consists of speech and nonspeech data where M denotes male and F denotes female respectively.

Table 1 and Table 2 represents MFCC feature extraction and LSF extraction the parameters respectively:

Analysis frame duration (ms)	Tw = 25
Analysis frame shift (ms)	Ts = 10
Pre-emphasis coefficient	alpha =0.97
Number of filterbank channels	M = 22
Number of cepstral coefficients	C = 18
Cepstral sine lifter parameter	L = 22
Lower frequency limit (Hz)	LF = 50
Upper frequency limit (Hz)	HF = 5000

Table 1

Analysis frame duration (ms)	Tw = 25
Analysis frame shift (ms)	Ts = 10
Number of poles	p = 12
Window size	wsize =551

Table 2

4.1. Performance Analysis Parameters

Following parameters are used for performance analysis of the speaker segmentation.

Precision is the ratio of correctly found changed (CFC) to the number of changes detected(DET) and recall is the ratio of correctly found changes to the actual speaker turns i.e ground truth(GT). The precision (PRC) and recall (RCL) rates given by:

$PRC = CFC/DET \frac{CFC}{DET}$, $RCL = CFC/GT$

where CFC denotes the number of correctly found changes and DET is the number of the detected speaker changes. Figure of merit can be represented by:

F1 =2(PRC*RCL) / PRC+RCL

Table 3 &Table 4 shows the results indicating the performance of speaker segmentation with MFCC and LSF features respectively

Database	Feature: MFCC					
	lambda=3, th2=0.7			lambda=1.5, th2=0.45		
	PRC	RCL	F1	PRC	RCL	F1
File1	1	0.67	0.80	0.66	0.8	0.72
File2	1	0.75	0.85	1	0.75	0.85
File3	0.66	0.5	0. 56	0.5	0.4	0.44
File4	1	0.85	0. 91	0.65	0.71	0.67
File5	1	0.75	0.85	0.75	0.75	0.83
Average	0.93	0.70	0.79	0.71	0.68	0.70

Table 3

Database	Feature: LSF					
	lambda=3, th2=0.7			lambda=1.5, th2=0.45		
	PRC	RCL	F1	PRC	RCL	F1
File1	1	0.67	0.80	0.66	0.8	0.72
File2	1	0.75	0.85	1	0.75	0.85
File3	1	0.75	0.85	0.66	0.50	0.56
F ile4	1	0.85	0.91	0.65	0.71	0.67
File5	1	0.75	0.85	0.75	0.75	0.83
Average	1	0.754	0.85	0.74	0.70	0.72

Table 4

5. Conclusion and Future Scope

Speaker segmentation is defined as the process by which a speech signal of long duration is partitioned into homogenous segments by detecting changes of speaker identity. Here speaker segmentation is carried out in this paper. The objective of the paper and the system architecture are also presented. Performance is analysed using two feature extraction techniques: MFCC and LSF. The precision, recall and figure of merit are calculated. The average figure of merits are 0.80 and 0.70 for MFCC and LSF respectively.

6. References

i. Miro, X. A., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. Audio, Speech, and Language Processing, IEEE Transactions on, 20(2), 356-370.

ii. Tranter, S. E., & Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. Audio, Speech, and Language Processing, IEEE Transactions on, 14(5), 1557-1565.

iii. Kotti, M., Moschou, V., & Kotropoulos, C. (2008). Speaker segmentation and clustering. Signal processing, 88(5), 1091-1124.

iv. Stafylakis, T., & Katsouros, V. (2011). A Review of Recent Advances in Speaker Diarization with Bayesian Methods. INTECH Open Access Publisher.

v. Hernawan, S. (2012, September). Speaker Diarization: Its Developments, Applications, and Challenges. In proceedings intl conf information system business competitiveness.

vi. Aronowitz, H. Speaker Diarization using Unsupervised Compensation of Within-Speaker Variability.

vii. Pérez-Cruz, F. (2008, July). Kullback-Leibler divergence estimation of continuous distributions. In Information Theory, 2008. ISIT 2008. IEEE International Symposium on (pp. 1666-1670). IEEE.

viii. Naresh, J., Holambe, R. S., & Basu, T. K. (2013, September). Unsupervised Speaker Segmentation and Clustering Using TESBCC and Pitch Based Features. In Computational Intelligence and Communication Networks (CICN), 2013 5th International Conference on (pp. 215-219). IEEE.

ix. L. Lu, H. Jiang, and H. J. Zhang (2001) , A robust audio classification and segmentation method, in Proc. 9th ACM Int. Conf. Multimedia, pp. 203–211.

x. R Huang, J. H. L. Hansen(May 2006) , Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora," Audio, speech, Language Process., vol. 1, no. 3, pp. 07919, IEEE Trans.