STUDY OF STOP LANDMARK DURATIONS FOR SPEAKER RECOGNITION

A dissertation
submitted in partial fulfillment of the
requirement for the degree of

Master of Technology

by

Welday Atsbaha Fisseha
(Roll No. 07307204)

under the supervision of

Prof. P. C. Pandey

Department of Electrical Engineering
Indian Institute of Technology Bombay
June 2009

W. A. Fisseha / Prof. P. C. Pandey (supervisor): “Study of stop landmark durations for speaker recognition”, *M.Tech. dissertation*, Department of Electrical Engineering, Indian Institute of Technology Bombay, June 2009.

ABSTRACT

The objective of this project is to investigate the use of stop landmark durations for improving speaker recognition. The variations of stop closure and burst durations across speakers are studied using variance tests. The results indicate that stop closure and burst durations may be used in combination with the spectral features for improving speaker recognition. Rate-of-rise (ROC) of mel-filtered squared magnitude spectrum is investigated for locating spectral transitions in the speech signal. Mel-filtering along the frequency axis improves landmark detection by enhancing the perceptually significant spectral transitions and smoothing harmonic structure and noise. Two automated methods for detecting stop closure, burst and frication offset landmarks are developed based on the ROC of mel-filtered spectrum. In the first method, the ROC peaks are selected using peak picking algorithm based on local threshold, and spectral slope, Wiener entropy and average magnitude spectrum are used as additional features to detect stop landmarks. In the second method, closure intervals in the speech signal are located based the product of Wiener entropy and log energy, and stop landmarks are detected by picking ROC peaks around the end points of the closures. The Wiener entropy and log energy are computed from the magnitude spectrum, while the spectral slope is computed from the mel-filtered squared magnitude spectrum. Landmark detection tests are carried out on VCV syllables and TIMIT sentences. Stop landmarks in VCVs are detected at rates of 53%, 75%, 90%, 95% and 97% respectively within 3, 5, 10, 15 and 30 ms of the manually labeled landmarks. The detection rates for TIMIT sentences are 52%, 69%, 83%, 87%, 90% and 93% within 3, 5, 10, 15, 20 and 30 ms of the manual landmarks, respectively.

Text-independent speaker recognition tests are conducted using Gaussian mixture modeling of closure and burst durations, and MFCC parameters. The performance of the duration features alone is not satisfactory, but an improvement of up to 4% is obtained using combination of MFCC and duration features. The results indicate that stop closure and burst durations convey speaker-dependent information and they could be potential candidates for improving speaker recognition.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor Prof. P. C. Pandey for the invaluable and unlimited guidance, motivation, and support he provided me throughout the project work. Besides the detailed discussions on the subject matter and research methods, his all-rounded and friendly advice, encouragement, and support was crucial for the completion of my work. I am also very grateful to Prof. Preeti Rao for her useful suggestions during my presentations.

I would like to sincerely thank Mr. A. R. Jayan who helped me a lot in collecting the resources required for my work, including his papers and speech database. I am also grateful to all my friends in the SPI Lab for their help during my stay.

I also sincerely thank the Board of Mekelle Institute of Technology (Ethiopia, Mekelle) and the Indian Council for Cultural Relations for jointly sponsoring my study programme in IIT Bombay. They covered all the necessary expenses, which helped me to finish my studies without difficulties.

Finally, I would like to thank my beloved family and friends for their support and encouragement during my M.Tech. study.

W. A. Fisseha

June 2009

CONTENTS

LIST OF SYMBOLS

Symbol	Explanation
α	Statistical significance level
λ	Speaker model
$\boldsymbol{\mu}_i$	Mean vector of GMM component *i*
$\mu_T(\mathbf{X})$	Mean impostor log likelihood scores
$\boldsymbol{\Sigma}_i$	**C**ovariance matrix of GMM component *i*
$\sigma_T(\mathbf{X})$	Standard deviation of impostor log likelihood scores.
a_k	Linear prediction coefficients
$a_n(k)$	First difference of mel-filtered squared magnitude spectrum along frequency axis
B	Linear bandwidth
B_L	Minimum mel-filter linear bandwidth
c_j	Cepstral coefficients
$D_n(k)$	First difference of mel-filtered squared magnitude spectrum along time axis
$E_L(n)$	log energy
$E_W(n)$	Wiener entropy
f	Linear-scale frequency
F	F-ratio value
f_C	Linear-scale center frequency
f_L	Linear-scale lower cut-off frequency
F_s	Sampling frequency
f_U	Linear-scale upper cut-off frequency
$\mathbf{H}(f)$	Mel-filter bank magnitude response
$H_k(m)$	Magnitude response of the k^{th} mel-filter
mel(f)	Mel-scale frequency
mel(f_C)	Mel-scale center frequency
mel(f_L)	Mel-scale lower cut-off frequency
mel(f_U)	Mel-scale upper cut-off frequency
MSS_a	Mean sum of squares due to factor A
MSS_ab	Mean sum of squares due to interaction effect of factors A and B

MSS_b	Mean sum of squares due to factor B
MSS_B	Mean sum of squares between groups
MSS_e	Mean error sum of squares
MSS_W	Mean sum of squares within a group
N	Number of FFT points
N_a	Number of levels of factor A
N_b	Number of levels of factor B
N_r	Number of repetitions (replicates)
N_T	Total number observations
$p(\mathbf{x})$	Probability of a feature vector **x**
$p(\mathbf{x}/\lambda)$	Likelihood of a feature vector **x** given speaker λ
$p(\mathbf{X}/\lambda)$	Likelihood of a sequence of feature vector **X** given speaker λ
S	Number of speakers
$s(n)$	Speech sample
$S(\mathbf{x}/\lambda)$	Un-normalized speaker score
$S_I(n)$	Sum of magnitude spectra
$S_n(k)$	Fourier transform of speech samples
SS_a	Sum of squares due to factor A
SS_{ab}	Sum of squares due to interaction effect of factors A and B
SS_b	Sum of squares due to factor B
SS_B	Sum of squares between groups
SS_e	Error sum of squares
SS_W	Sum of squares within a group
$S_T(\mathbf{x}/\lambda)$	Test normalized speaker score
T	Number of training or testing feature vectors
w_i	Mixture weight of GMM component i
x	Feature vector
X	Sequence of feature vectors
$X_S(n)$	Product of Wiener entropy and log energy
$x_{s,t}$	Stop closure duration for speaker s at the t^{th} trial
$\bar{x}_s$	Mean stop closure duration for speaker s
$\bar{\bar{x}}$	Mean of speaker-wise means

LIST OF ABBREVIATIONS

Abbreviation	Explanation
ANOVA	Analysis of variance
ASI	Automatic speaker identification
ASV	Automatic speaker verification
CMS	Cepstral mean subtraction
CMVS	Cepstral mean and variance subtraction
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DTW	Dynamic time warping
EM	Expectation maximization
FFT	Fast Fourier transform
GMM	Gaussian mixture model
HMM	Hidden Markov model
HNM	Harmonic plus noise model
LMD1	Landmark detector – 1 (based on rate-of-change of mel-filtered magnitude spectrum and voicing detection)
LMD2	Landmark detector – 2 (based on rate-of-change of mel-filtered magnitude spectru and silence detection)
LP	Linear prediction
LPC	Linear prediction coefficients
LPCC	Linear prediction cepstral coefficients
MFCC	Mel-frequency cepstral coefficients
PLP	Perceptual linear prediction
PLPC	Perceptual linear prediction cepstral coefficients
RASTA	Relative spectra
ROC	Rate of change
SWT	Stationary wavelet transform
TD	Text-dependent
TI	Text-independent
VOT	Voicing onset time
VQ	Vector quantization

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Problem Overview

Voices of different speakers have different acoustic and perceptual characteristics that enable us to distinguish between speakers just by listening to their voices. Speaker recognition is the task of determining the speaker of a sample utterance using speaker-dependent information extracted from it [1], [2]. The variations in the characteristics of speech utterances of different speakers result from differences in the anatomy of the speech production system and learned speaking habits of the speakers [1] - [5]. Anatomical differences are related to the fixed structural differences in the shape and size of the speech production organs, mainly vocal tract and vocal folds. Learned speaking habits, on the other hand, refer to the way individuals use their speech mechanism, the movements of the organs and their language usage [2]. Speaker recognition systems use both physiological and learned characteristics of speakers to extract speaker-dependent features such as mel-frequency cepstral coefficients (MFCC) or linear prediction cepstral coefficients (LPCC) and their dynamics, and pitch and timing patterns [1], [2], [6].

Speaker recognition is a pattern matching problem which involves two phases: training and testing [2], [7]. In the training phase, speaker discriminating features are extracted from sample utterances of a known speaker and a model representing the speaker is generated. Speaker models can be templates (deterministic) such as vector quantization (VQ) and dynamic time warping (DTW), or stochastic such as hidden Markov models (HMM) and Gaussian mixture models (GMM). In the testing phase, speaker-dependent feature vectors are extracted from the unknown test utterance and compared with existing speaker models to recognize the speaker. Comparison is based on distance or similarity measures for template models, and probability measures for stochastic models.

The main difficulty in speaker recognition is the lack of explicitly measurable attributes of speech signal that can effectively discriminate among speakers. Therefore, identifying, extracting, and efficiently modeling the embedded speaker-dependent features of speech signal is crucial to reliably distinguish speakers. Existing speaker recognition systems commonly use linear prediction and cepstral analysis to extract vocal tract features such as LPCCs and MFCCs. These systems perform well with normal speech under favorable conditions, but their performance degrades under practical environmental conditions and various speaker efforts such as mimicking.

Therefore, studying additional features of the speech signal, and suitable front-end processing and extraction techniques will be important to improve the performance of speaker recognition systems.

1.2 Project Objective

The objective of this project is to investigate the use of stop duration features for improving speaker recognition and to develop a landmark detection method for extracting the duration parameters. Towards this end, the variation of stop closure and frication burst durations within and across speakers is studied on recorded speech using analysis of variance. The effects of speaking rate and context variability on stop durations are also studied using two-way analysis of variance tests. A method for detecting stop landmarks and measuring their durations is developed based on rate-of-change of mel-filtered magnitude spectrum, spectral slope, spectral flatness measure, and energy parameters. The performance of the method is evaluated on VCV syllables and TIMIT sentences. A text-independent speaker recognition system based on Gaussian mixture modeling of stop duration and MFCC features is investigated and evaluated using TIMIT database.

1.3 Thesis Outline

The next chapter gives a review of the basic steps in speaker recognition. The various classifications, applications, and the basic components of speaker recognition such as feature extraction, feature selection, speaker models, pattern matching and moralization techniques are explained. In Chapter 3, speaker recognition features such as linear prediction cepstral coefficients (LPCC), mel-frequency cepstral coefficients (MFCC), harmonic plus noise model (HNM) parameters, and high level features are discussed. Statistical analysis of variation of stop closure and burst durations across speakers is given in Chapter 4. In Chapter 5, a method for detecting spectral transitions based on the rate-of-change (ROC) of mel-filtered magnitude spectrum is investigated. A stop landmark detector based on ROC, spectral flatness measure, spectral slope and signal energy is discussed. Evaluation results on VCV syllables and TIMIT sentences are presented. Chapter 6 discusses GMM speaker recognition system using stop closure and burst durations, and MFCC features. Speaker recognition results on TIMIT database are presented. A summary of the work done and conclusion are given in Chapter 7. Detailed results obtained from the analysis of variance of stop closure duration at different speaking rates are given in Appendix A.

Chapter 2

SPEAKER RECOGNITION

2.1 Introduction

There are two types of speaker recognition tasks: automatic speaker verification (ASV) and automatic speaker identification (ASI) [1], [2]. Fig. 2.1 gives a block diagram representation of the verification and identification processes. In ASV, the objective is to determine if a claimant is the person s/he claims to be [1], [2], [5]. The system compares features from a sample utterance of the claimant with the speaker model representing the claimed identity and the decision is to accept or reject the person by comparing the similarity measure with a threshold. In ASI, on the other hand, the task is to determine the speaker of a given speech sample [2], [3]. If it is known that the model of the unknown speaker exists in a given set of reference models, the test is called closed-set identification, and the goal is to identify the talker of the sample utterance from the set of known speakers, also called customers. The unknown feature vector is compared with each of the models to determine the best match. When it is possible that the model of the unknown speaker does not exist in the given reference set, the test is called open-set identification. In this case, the recognition process combines the tasks of verification and closed- set identification. The system has to determine whether the unknown speaker belongs to the given set of customers (verification), and identify which particular person s/he is [1], [8]. The performance of ASI system degrades as the number of reference models increases, while ASV can perform well independent of the population size [3].

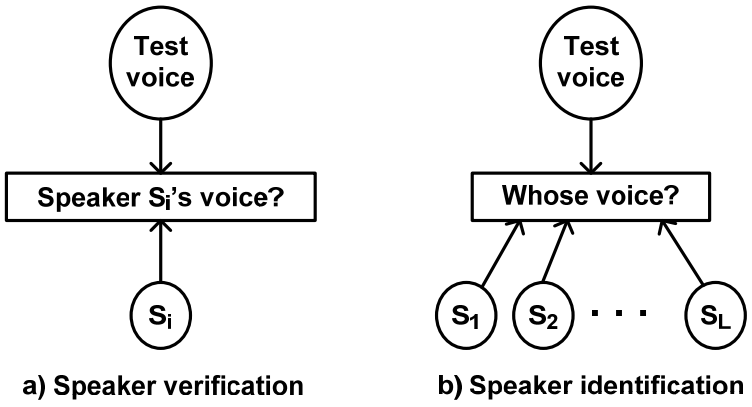

Fig. 2.1 Speaker verification and identification [9]

Based on whether the training and test utterances are the same or not, speaker recognition systems are classified as text-dependent (TD) and text-independent (TI) [1], [2]. In TD system, the same utterance is used for training and testing. This method is used for high security applications with cooperative users. The use of the same text for training and testing simplifies system complexity and improves the recognition accuracy. In situations where users are not cooperative or when the speaker recognition is to be done without the knowledge of the target person, TI system is used, in which the training and test utterances are different. Compared to TD, TI systems are more complex and require more training data to achieve good performance due to the additional task of overcoming text-dependent variations [3].

Speaker recognition has been extensively used for security and authentication applications such as [1], [2], [3]

- verifying the identity of individuals prior to admission to secure area,
- remote authentication of customers for telephone banking, telephone credit cards and access to secure information system,
- identifying talkers in audio conferences,
- alerting speech recognition systems of change of speakers, and
- checking if a user is already enrolled in a system.

Speaker recognition is also used to identify the gender or accent of a speaker, language spoken [3], and in forensic applications to verify or identify a criminal from a given criminal voice sample [5]. Use of biometric recognition for security applications has some advantages over the classical methods which depend on something you own such as key and card, or something you know such as password and code, as these can be stolen, lost or forgotten [3], [5]. Compared to other biometric traits such as fingerprint, iris, and face, speech offers a greater flexibility and different levels of recognition can be used for different applications [5]. Speaker recognition systems can ask the user to speak in a particular way, and the system can use a variety of parameters to differentiate speakers. Besides, speech data can be collected easily without the knowledge of the speakers or over the telephone.

The performance of a speaker recognition system gets affected by noise and various mismatch conditions which can lead to two types of recognition errors [3]: false acceptance (FA or Type I error) and false rejection (FR or Type II error). FA occurs when the system incorrectly recognizes an impostor as the claimed speaker while FR occurs when the system rejects a true claimant in ASV or incorrectly finds no match in ASI. The trade-off between FA and FR is used to determine the decision threshold, which is chosen to minimize the overall risk. Usually FA errors are more risky, hence low thresholds are preferred to minimize false acceptance of malicious impostors.

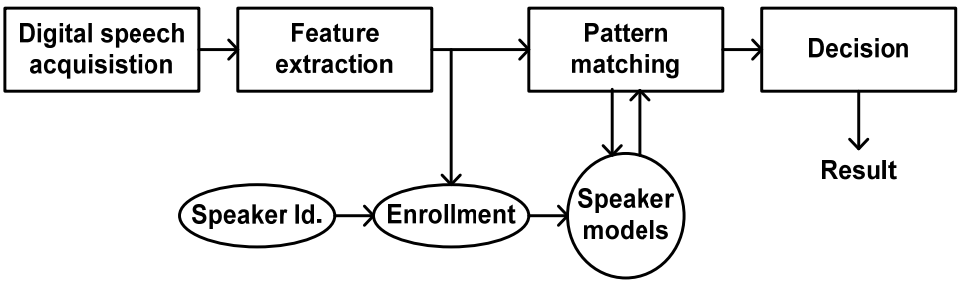

Fig. 2.2 Block diagram of a speaker recognition system [2]

There are a number of factors that cause recognition errors, which can be grouped under algorithmic and non-algorithmic factors. Algorithmic factors include among others the use of insufficient or poor speaker-dependent parameters and less fitting models. The non-algorithmic factors are intra-speaker variability, noise, and mismatches in recording and transmission environments. Intra-speaker variations result from variations in speaker conditions due to sickness, emotional state, aging, etc. Noise and channel mismatch errors can be minimized by applying normalization and adaptations techniques on the features and the recognitions scores [10] - [14].

2.2 Feature Extraction/Selection

Speaker recognition involves the following basic steps [2], [15]: speech data acquisition, feature extraction and selection, pattern matching (modeling) and classification, decision logic, and enrolment to generate speaker models. Figure 2.2 shows a general block diagram of a speaker recognition system. The analog speech signal obtained from microphone is converted into digital form through sampling and quantization. Filtering, silence removal and channel compensation techniques are applied before further processing [2], [8], [15]. The resulting signal is split into frames of 10-30 ms and a speaker-dependent feature vector is extracted for each frame to create speaker models. In the testing phase, feature vectors computed from the test utterance are compared with the reference speaker models and recognition decision is made based on the matching score.

The performance of a speaker recognition system critically depends on the extraction of features which are effective in discriminating speakers. A feature is effective if it has low intra-speaker and high inter-speaker variability [2], [3]. A careful selection of independent parameters that are capable of efficiently representing the speaker-dependent features permits the use of simplified mathematical tools for pattern matching and classification. Features for speaker recognition should [1], [3]

- have low intra-speaker and high inter-speaker variances,
- be efficient in representing the speaker-dependent information,

- be easy to extract,
- be stable over time,
- be frequently and naturally occurring in speech,
- be robust to noise, and
- be less susceptible to mimicry.

One measure of the effectiveness of recognition parameters is comparing the inter-speaker and intra-speaker variances using F-ratio defined as [1], [2], [3]

$$F = \frac{\text{variance of speaker means}}{\text{mean of speaker variances}} \tag{2.1}$$

The F-ratio value will be large if the values of the parameter are widely spread for different speakers and less variable within a speaker. F-ratio fails when the speaker means are by chance the same in which case the difference in mean becomes zero and the parameters would be regarded as unimportant. F-ratios are most useful for eliminating poor features rather than selecting the best [3].The best way to evaluate the effectiveness of a feature is in terms of the probability of error in recognizing a speaker [2], [3]. However, probability of error also depends on the model used; hence computation becomes complex and time-consuming. F-ratio does not take into account the correlations among parameters which might result in the selection of redundant parameters [2], [3]. A generalization of the F-ratio measure, called divergence is used for estimating the combined effectiveness of a set of features in discriminating among speakers. It is computed as [2]

$$D = \left\langle \left[\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right] \boldsymbol{\Sigma}^{-1} \left[\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right]^T \right\rangle_{i,j} \tag{2.2}$$

where $\boldsymbol{\mu}_i$ is the mean feature vector for speaker i, $\langle\ \rangle_{i,j}$ represents average over all speakers, $1 \leq i,\ j \leq N$, and $\boldsymbol{\Sigma}^{-1}$ is the intra-speaker inverse covariance matrix. Divergence is a measure of dissimilarity among the feature vectors of speakers, including the interdependence between individual features.

2.3 Speaker Modeling

After extracting speaker-dependent features, the next step in speaker recognition is pattern matching (modeling and classification). During enrolment, a model of the speaker is created from the training features vectors, and stored. In the testing stage, a match score which is a measure of similarity between the test feature vector and reference model is computed. Establishing effective and computationally efficient speaker models is a difficult problem in speaker recognition, because speech signal and the environmental conditions which affect it are highly variable and unpredictable.

There are two categories of models used in speaker recognition [1], [2], [9]: template models and stochastic models. A template model consists of a sequence of feature vectors extracted from the training utterance. The test feature vector is assumed to be a distorted copy of the template vector, and they are aligned using a warping function that minimizes the distance between them [2], [3]. Examples of template models include dynamic time warping (DTW) and vector quantization (VQ). Stochastic models represent a speaker in terms of probability distributions estimated from the training data. The match score is computed as the likelihood or conditional probability of the test feature vectors given the speaker model. Hidden Markov models (HMM) and Gaussian mixture models (GMM) are among the common stochastic models. DTW outperforms VQ and the stochastic models when sparse training data are used. However, the computational complexity in DTW increases as the square of the template duration which makes it difficult to use for long training utterances [3]. VQ performs better than HMM and GMM under moderate amount of training data; however, GMM is generally preferred because it is more robust to noise, and is capable of modeling the distribution of unconstrained speech [12], [16], [17].

2.3.1 Template Models

Template models are used for text-dependent speaker recognition. Speakers are modeled in a non-parametric manner by sequences of feature vectors extracted from repeated utterances of the same phrase. The pattern matching algorithm tries to align similar features of the model and the observation sequence such that the separation between the two templates is minimized. Euclidian and Mahalanobis distances are commonly used to measure the similarity between the templates. The Euclidian distance between two feature vectors $\mathbf{x}$ and $\mathbf{y}$ is given by [2], [3]

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^T} \tag{2.3}$$

The Mahalanobis distance between the two vectors is defined as

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})^T} \tag{2.4}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the two vectors [2], [3], given as

$$\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})(\mathbf{y}-\boldsymbol{\mu}_\mathbf{y})^T] \tag{2.5}$$

where $E[.]$ represents expectation (first moment) operation, and $\boldsymbol{\mu}_\mathbf{x}$ and $\boldsymbol{\mu}_\mathbf{y}$ are the mean vectors of $\mathbf{x}$ and $\mathbf{y}$ respectively. The multiplication factor $\boldsymbol{\Sigma}^{-1}$ in (2.4) gives greater weights to the vector parameters which are more effective in distinguishing speakers. When $\boldsymbol{\Sigma}^{-1}$ is identity matrix, Mahalanobis distance reduces to Euclidian distance. Euclidian distance is the optimal measure when the parameters are mutually independent and have equal variances, but Mahalanobis distance has the advantage that it is invariant to non-singular transformations [2].

Template models that have been extensively used since the early investigations of speech and speaker recognition are dynamic time warping (DTW) and Vector quantization (VQ) [18]. DTW involves alignment and distance computation, and is used for text-dependent speaker recognition. It addresses the problem of speaking rate variability by warping the time scale of the reference template in order that similar events in the two templates occur at the same time [3]. Given variable length reference templates $\{\mathbf{R_1}, \mathbf{R_2}... \mathbf{R_T}\}$ and a test template **T,** DTW finds a warping function $n = w(m)$ which maps the principal time axis of **T** to the time axis of **R**. In comparing multiframe templates, Bellman optimality principle is used to determine the best path among numerous possibilities. The warping function is computed to minimize the total distance measure between the test and reference templates, defined by

$$D = \min_{w(n)} \left[\sum_{n=1}^{T} d(T(n), R(w(n))) \right] \tag{2.6}$$

where each d term is a frame distance between the n^{th} test frame and the $w(n)^{th}$ reference frame, and D is the minimum overall distance measure corresponding to the best path [3].

The warping function must satisfy the end point constraints and monotonicity. Given a feature vector of a test utterance, DTW determines the optimum alignment warping function and computes the distance or distortion of the feature vector from all the reference templates to find the best match. In speaker identification, the decision rule is to select the reference template **R*** with the smallest alignment distortion or distance. For verification, the distortion between the claimant's feature vector and the model of the claimed identity is evaluated and scored against threshold to make a decision. DTW works well when the number and duration of reference templates is small. As the size of the population increases, computation of the warping function and comparison of the test template with all the reference templates becomes complex.

In text-independent speaker recognition, a set of feature vectors extracted from short training utterances can be used directly as the speaker model [8]. However, such a direct representation becomes difficult when a large number of feature vectors are used as memory requirement and computational complexity makes the system costly. In addition, there is a large dependence between successive feature vectors of an utterance, and hence redundancy in representation. Vector quantization model is a reduced representation of the speaker's feature vectors, which is commonly used for text-independent speaker recognition. The space of the feature vectors extracted from the person's utterances is split into a finite number of regions and each region is represented by its center, called a codeword. Each speaker is represented by a unique collection of codewords, called a codebook, generated from the training data using standard clustering algorithms. One method for designing a codebook is the K-means clustering

algorithm, an iterative procedure that converges to a final codebook by minimizing the average distortion across the training set [19]. The algorithm has the following steps:

1. The number of required clusters, also called the size of the codebook, *K*, is fixed.
2. Initial codewords, also called cluster centroids, are randomly selected from the input vectors.
3. The Mahalanobis or Euclidean distance between each input feature vector and each centroid is determined. Each vector joins the cluster of the centroid that gives the minimum distance.
4. New centroids are computed for the clusters formed by taking the average of each cluster.
5. Steps 3 and 4 are repeated until there is no change in the centroids or the change is small.

In this way, the entire space of feature vectors is partitioned into *K* regions. Any input feature vector is classified as belonging to one of the regions. The centroids form the codewords of the codebook that represents the reference model of the particular speaker. One codebook of *K* codewords is created for each of the enrolled speakers during the training phase.

During the testing, a set of *T* feature vectors are extracted from the test utterance and each feature vector is quantized using the codebooks of all reference models in the database. Next, the average quantization distortion Q_s for each reference speaker *s* is computed as

$$Q_s = \frac{1}{T}\sum_{n=1}^{T} d\left(\mathbf{x}(n), \mathbf{v}_s^*\right) \tag{2.7}$$

where $\mathbf{x}(n)$ is the input feature vector, $\mathbf{v}_s^*$ is a codeword vector, from the codebook of the reference speaker, which is closest to $\mathbf{x}(n)$, and $d(\mathbf{x}(n), \mathbf{v}_s^*)$ is the distortion measure between the two vectors. Finally, the speaker with the minimum average distortion is identified as the claimed speaker.

VQ performs well in both text-dependent and text-independent recognition systems with relatively short utterances [3]. Compared to DTW, VQ requires less computation and storage. The clustering algorithm used to create the codebook averages out temporal information from the codewords [2]. This simplifies the system by avoiding the need for time alignment, but it may also degrade the system performance by removing speaker-dependent temporal information.

2.3.2 Stochastic Models

Template models are deterministic pattern matching approaches that have been used dominantly in the early works of speaker recognition. These methods are intuitively reasonable and perform well with small amount of training data [2], [3]. However, currently they are being replaced by stochastic models which provide more flexibility and theoretically meaningful likelihood scores. In stochastic models, the sequence of feature vectors extracted from the training data are assumed

as random processes, and a parametric or non-parametric density function is estimated from the feature vectors to represent each speaker. During testing, the likelihood or conditional probability of the test feature vectors given the reference model is computed to make recognition decision. In a closed-set identification, the decision rule is to choose the reference speaker which has the greatest probability of generating the test feature vectors. Suppose λ_s is the stochastic model for speaker *s* in a system with *S* speaker models; each speaker is represented by one stochastic model generated from the training data. Let **X** be a sequence of feature vectors obtained from a test voice sample Speaker s^* is identified as the target speaker if

$$p(\mathbf{X}/\lambda_{s^*}) \geq p(\mathbf{X}/\lambda_s),\ 1 \leq s, s^* \leq S,\ s^* \neq s \tag{2.8}$$

where $p(.)$ represents probability. For verification, the computed likelihood value is compared with a threshold to make a decision. The main problem in stochastic models is the estimation of appropriate joint probability density function for the sequence of feature vectors and computation of the test likelihood. Usually, the likelihood computation is simplified by assuming successive feature vectors to be independent [2].

Stochastic models provide a better model of acoustic events and allow the application of newly developed noise and channel adaptation methods [12]. There are a number of stochastic models used for modeling random observation sequences, of which Gaussian mixture models GMM [12], [16] and hidden Markov models (HMM) [2], [18, [20] are often used for speech and speaker recognition. HMMs are finite state machines where the observation vectors in each state are probabilistic functions of the state. They are used for both text-dependent and independent speaker recognition [2], [8], [21]. An HMM model is specified by [8]

1. the number of states in the model, *K*;
2. the number of distinct observation symbols per state, L;
3. the state transition probability distribution, **A**;
4. the observation symbols probability distribution in each state, **B**; and
5. the initial state distribution, $\boldsymbol{\pi}$.

During training, the HMM parameters are estimated from the feature vectors of the training data. This involves computation of the state-transition and the observation symbol probabilities for each state using established algorithms such as Viterbi segmentation and forward-backward algorithm [15], [18]. In the testing stage, the likelihood of the test feature vectors given the speaker model is computed using Viterbi algorithm to make verification or identification. The performance of HMM is in most cases (except with sparse training data) comparable to that of VQ, and the stochastic Markovain transitions between states in HMM provide the advantage of representing temporal variations in the parameters of the speech signal [2], [8], [18].

GMM is used for text-independent speaker recognition and provides high performance for real-time applications [12], [16]. In this case, the distribution of the feature vectors extracted from a speaker's utterance is modeled as a weighted sum of uni-modal multivariate Gaussian density functions [12]. The speaker model is represented by the mean vectors, covariance matrices and mixture weights of the component functions, which are estimated from the speaker's training feature vectors using expectation maximization (EM) algorithm. During testing, the likelihood of the test feature vector given the speaker model is computed, and the decision is made based on comparing with a threshold for verification and maximum likelihood for identification. We have used GMM for our speaker recognition and it will be discussed further in the next section.

2.4 Gaussian Mixture Models (GMM)

In GMM, the conditional probability density function of a *D*-dimensional observation vector **x,** given a model λ, denoted as $p(\mathbf{x}/\lambda)$, is approximated by a weighted sum of multivariate Gaussian density functions. This is mathematically defined as [12]

$$p(\mathbf{x}/\lambda) = \sum_{i=1}^{M} w_i p_i(\mathbf{x}) \tag{2.9}$$

where M is the number of mixtures, and $p_i(\mathbf{x})$ and w_i, $1 \leq i \leq M$, are the component Gaussian functions and their mixture weights, respectively. Each component Gaussian function $p_i(\mathbf{x})$ is given by

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} \mid \mathbf{\Sigma}_i \mid^{1/2}} e^{-1/2(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)} \tag{2.10}$$

where $\boldsymbol{\mu}_i$ is a *D*-dimensional mean vector and $\mathbf{\Sigma}_i$ a $D \times D$ covariance matrix. The mixture weights (priors) satisfy the constraint

$$\sum_{i=1}^{M} w_i = 1 \tag{2.11}$$

The GMM model λ for each speaker is represented in terms of the means, covariances and mixture weights of the component functions as

$$\lambda = \{w_i, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\}, 1 \leq i \leq M \tag{2.12}$$

During training, the GMM parameters are estimated from the training feature vectors using expectation maximization (EM) algorithm [12]. The EM algorithm iteratively modifies the GMM parameters such that the likelihood of the feature vectors given the model increases monotonically. Given a sequence of T training feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, the likelihood score $p(\mathbf{X}/\lambda)$ is computed as

$$p(\mathbf{X}/\boldsymbol{\lambda}) = \prod_{n=1}^{T} p(\mathbf{x}_n/\boldsymbol{\lambda}) \tag{2.13}$$

The product comes from the assumption that the individual feature vectors are statistically independent. The objective in EM estimation is to find new parameters $\hat{\boldsymbol{\lambda}}$ such that the likelihood of the training vectors is maximized, i.e.,

$$p(\mathbf{X}/\hat{\lambda}) \geq p(\mathbf{X}/\lambda) \tag{2.14}$$

Starting with an initial model, the EM computes the new model parameters using a two step iterative process. In the first step, called the expectation-step, the a posterior probabilities of each component *i* to all the training vectors to is computed using [12]

$$p(i/\mathbf{x}_n, \boldsymbol{\lambda}) = \frac{w_i p_i(\mathbf{x}_n)}{\sum_{k=1}^{M} w_k p_k(\mathbf{x}_n)}, 1 \leq n \leq T \tag{2.15}$$

In the second step, called the maximization-step, new component weights, means and covariances are computed from the initial parameters and a posterior probabilities, as follows:

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{n=1}^{T} p(i/\mathbf{x}_n, \boldsymbol{\lambda})\mathbf{x}_n}{\sum_{n=1}^{T} p(i/\mathbf{x}_n, \boldsymbol{\lambda})} \tag{2.16}$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{\sum_{n=1}^{T} p(i/\mathbf{x}_n, \boldsymbol{\lambda})(\mathbf{x}_n - \boldsymbol{\mu}_i)(\mathbf{x}_n - \boldsymbol{\mu}_i)^T}{\sum_{n=1}^{T} p(i/\mathbf{x}_n, \boldsymbol{\lambda})} \tag{2.17}$$

$$\hat{p}_i = \frac{1}{T}\sum_{n=1}^{T} p(i/\mathbf{x}_n, \boldsymbol{\lambda}) \tag{2.18}$$

The algorithm is terminated when the likelihood increment is below a certain threshold or by fixing the maximum number of iterations. The initial model is often selected using K-means clustering or vector quantization, and the number of mixtures depends on the amount of training data.

In the testing stage, the log likelihood of the test feature vector given a speaker model is computed for all the speakers in the database and the speaker with the maximum likelihood is recognized as the speaker of the test utterance. For a sequence of test feature vector $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, the log likelihood given a model λ_k is calculated as

$$\log p(\mathbf{X}/\boldsymbol{\lambda}_k) = \frac{1}{T}\sum_{n=1}^{T} \log p(\mathbf{x}_n/\lambda_k) \tag{2.19}$$

This expression is obtained from the a posterior probability by assuming equally likely speakers and independent feature vectors. Given a set of S speaker models $\lambda_1, \lambda_2, \ldots, \lambda_S$, the decision rule is to choose the speaker λ_k such that

$$\log p(\mathbf{X}/\lambda_k) \geq \log p(\mathbf{X}/\lambda_i),\ 1 \leq i, k \leq S,\ k \neq i \tag{2.20}$$

In order to compensate for score variability due to difference in test utterances, the distribution of scores is often normalized using T-norm [10] as described in the next section. The test feature **X** is scored against a set of imposter models, and the mean and variance of the imposter log likelihood scores is computed. The test normalized log likelihood score for a given speaker λ_k, $S_T(\mathbf{X}/\lambda_k)$, is obtained as follows:

$$S_T(\mathbf{X}/\lambda_k) = \frac{\log p(\mathbf{X}/\lambda_k) - \mu_T(\mathbf{X})}{\sigma_T(\mathbf{X})} \tag{2.21}$$

where $\mu_T(\mathbf{X})$ is the mean and $\sigma_T(\mathbf{X})$ is square root of variance of the imposter scores.

2.5 Feature and Score Normalization

The performance of a speaker recognition system is affected by the variability of speech between training and testing sessions. This variability is caused by mismatches between training and testing environments such as differences in the recording and transmission media, noise, and changes in the person's speech due to health problem or aging [10] - [14]. The primary approach to tackle this problem is to use recognition features and modeling techniques which are robust to the various mismatches. In addition, depending on the type of mismatch, various enhancement and normalization methods are used to improve recognition performance. These normalization techniques are categorized as feature normalization and score normalization.

In feature normalization, the individual feature vectors or their distribution is modified. Some of the feature normalization methods include spectral subtraction [22], cepstral mean subtraction (CMS) and ceptral mean and variance subtraction (CMVS) [14], relative spectra (RASTA) filtering [11] and feature warping [23]. In spectral subtraction [14], noise is assumed to be additive, uncorrelated with the speech signal and slowly varying, and an estimate of the noise obtained from non-speech interval is subtracted from the speech signal in the power spectrum domain. CMS [14] is used to reduce linear channel effects and involves subtracting a long-term average of cesptral features from each feature vector. In CMVS [13], mean and standard deviation of the feature vectors are estimated over specific segments and the distribution of the feature vectors is normalized by subtracting the mean and dividing by the standard deviation to minimize additive noise. RASTA processing [11] is a kind of modulation spectrum technique applied to reduce non-speech spectral components which vary faster or slower than the changes in the speech signal. In feature warping [23], the distribution of the individual feature streams is mapped

to a standardized distribution (typically normal distribution) to reduce additive noise and linear channel effects.

In speaker verification, score normalization is applied to compensate for recognition score variations among speakers and test utterances so that a global decision threshold can be used. Score normalization methods include model normalization, cohort normalization [10], Z-norm [13], T-norm [10] and H-norm [17], etc. A background model is a speaker-independent model generated using a collection of training utterances from different speakers. The normalization is carried out by dividing the likelihood score of the target speaker by the likelihood score of the background model [17]. In cohort normalization [10], instead of a single model, a set of imposter speakers who have similar characteristics to the target speaker are used to scale the likelihood score of each speaker.

Zero normalization (Z-norm) is used to reduce score variability due to different test utterances [13]. In this case, a speaker model λ_k is scored against a number of imposter utterances and mean and variance of the scores is estimated. During testing, the speaker's score against the test feature vectors **X,** $S(\mathbf{X}/\lambda_k)$ is computed and it is normalized as

$$S_Z(\mathbf{X}/\lambda_k) = \frac{S(\mathbf{X}/\lambda_k) - \mu_k}{\sigma_k} \tag{2.22}$$

where $S_Z(\mathbf{X}/\lambda_k)$ is the normalized score, and μ_k and σ_k are the speaker-specific mean and square root of variance of the imposter scores [13]. Test normalization (T-norm) is similar to Z-norm, but the normalization mean and variance are estimated from scoring the test utterance against a number of imposter speakers [10]. Handset normalization (H-norm) [17] is also a kind of Z-norm used to compensate for handset variability, in which the normalization parameters are estimated from handset-dependent imposter utterances.

Chapter 3

FEATURES FOR SPEAKER RECOGNITION

3.1 Introduction

The main task in speaker recognition is to identify, extract, and characterize the speaker-dependent information in speech signal for recognizing speakers. Extracting and selecting parameters which can effectively and efficiently represent the speaker distinguishing characteristics of speech is crucial for the success of speaker recognition systems. The desirable characteristics of speaker recognition parameters are mentioned in section 2.2. Particularly, independent parameters that have high inter-speaker and low intra-speaker variances are preferred.

There are two sources of speaker-dependent characteristics of speech: physiological and learned speaking styles [1] - [5]. Physiological differences are related to the geometric characteristics of the speech production system, mainly the shape and size of the vocal tract and vocal folds. These physical variations are manifested in the acoustic and perceptual features of the speech signal. Vocal tract characteristics refer to the formant frequencies and formant bandwidths which are dependent on the length, area and shape of the vocal tract. The length, tension and mass of the vocal folds determine the fundamental frequency [2]. Learned speaking habits, on the other hand, refer to the way individuals use their speech mechanism, the movements of the organs and language usage. Some of the learned speaking habits used for speaker recognition include speaking and pause rate, pitch and timing patterns, idiosyncratic word/phrase usage, idiosyncratic pronunciations, etc [1], [2], [3].

The most successfully used speaker recognition parameters are linear prediction cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC). These parameters are extracted using linear predictive coding analysis and cepstral analysis. The following sections give a review of the commonly used speaker recognition features and their extraction techniques.

3.2 Linear Prediction (LP) Analysis

Linear prediction is a powerful speech analysis technique used for estimating vocal tract and excitation parameters [2], [3], [24], [25]. It is based on the discrete-time model of the speech production system in which the vocal tract is assumed as a linear, time-varying filter excited by a quasi-periodic impulse train for voiced speech or random noise for unvoiced speech. In most of

the applications of LP, the vocal tract is modeled as an all-pole filter, defined in the frequency domain by [3]

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.1}$$

where G is a scaling factor (gain), p is the prediction order, and $\{a_k\}_{k=1,\ldots,p}$ are the prediction coefficients. In the time domain, the output speech sample $s(n)$ is given as the linear combination of p previous output speech samples and a scaled excitation:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{3.2}$$

where $u(n)$ is the excitation and $\{s(n-k)\}_{k=1,\ldots,p}$ are the past outputs. Since the input sequence, $u(n)$, is generally unknown [1], the prediction equation is simplified to

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{3.3}$$

The difference between the actual and the predicted values is called prediction or residual error $e(n)$, and is calculated as

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{3.4}$$

It can be observed that the error signal is directly related to the source excitation which carries speaker-dependent characteristics; and hence it is an important parameter for speaker recognition.

The prediction coefficients $\{a_k\}$ are chosen such that the prediction error is minimized in the mean squared sense. One reason for using mean squared error prediction is that it allows simple and efficient computation of the gain and prediction coefficients. Depending on the range of the interval considered, the autocorrelation, covariance or lattice methods can be used to compute the LP coefficients from the sampled speech signal [2], [3], [24], [25]. The all-pole model of speech production is appropriate for non-nasalized voiced sounds [3]. For nasals and fricative sounds, the model needs to include both poles and zeros to account for the anti-resonances introduced during nasalization and frication (trapping of energy at the mouth end). However, an all-pole filter with high prediction order provides a good approximation of the production of almost all speech sounds.

The LP coefficients convey vocal tract characteristic that are inherent to each speaker [2], [3], [26]. They are suitable for speaker recognition because they contain combined information about formant frequencies, formant bandwidths and the glottal wave. LP coefficients provide useful information about the vocal tract only as a set, and cannot be individually smoothed across analysis frames [2]. Therefore, they are often transformed into acoustically or perceptually

meaningful parameters such as reflection coefficients, log area ratios, linear spectral pair frequencies, and cepstral coefficients (LPCC). Description and comparison of these parameters is given in [2]. Among these parameters, LPCCs have been most effective for speaker recognition [1], [16], [28]. Figure 3.1 shows the block diagram of LPCC extractor.

The LP cepstral coefficients, defined as the inverse Fourier transform of the logarithm of the overall LPC system H(z), are computed from the LP coefficients using the following recursive formula [1], [16]:

$$\begin{aligned} c_1 &= a_{1,} \\ c_j &= a_j + \sum_{k=1}^{n-1}\left(1-\frac{k}{j}\right)a_k c_{j-k},\ 1 \le j \le p, \\ c_j &= \sum_{k=1}^{j-1}\left(1-\frac{k}{j}\right)a_k c_{j-k},\ j > p \end{aligned} \tag{3.5}$$

where a_j's and c_j's are the j^{th} order LP and LP cepstral coefficients respectively. A modified LP analysis called perceptual linear prediction (PLP) is used for improved speaker recognition, but at the cost of additional computational complexity [24], [28], [29]. In modeling the vocal system, PLP takes into account the human perception of sound such that greater emphasis is given to the

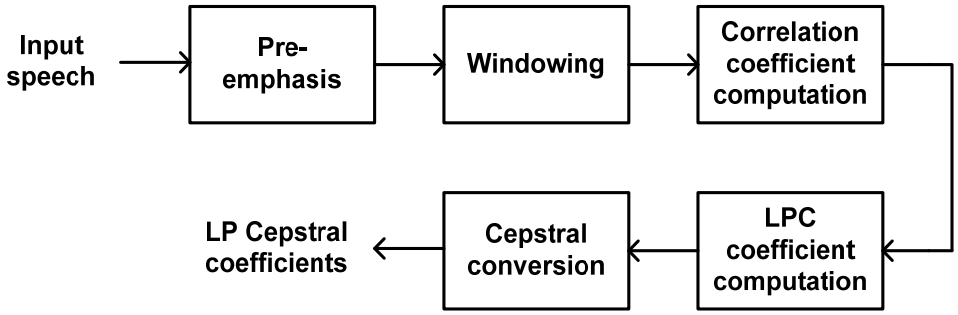

Fig. 3.1 LPCC extraction [16]

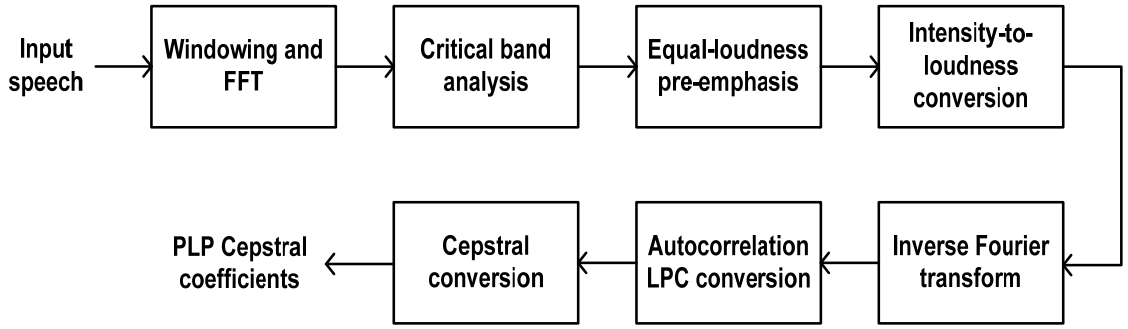

Fig. 3.2 Extraction of perceptual linear prediction cepstral coefficients (PLPC) [26]

perceptually important portions of the spectrum. As shown in Fig. 3.2, computation of PLP cepstral coefficients involves linear prediction cepstral extraction preceded by the following three steps [26]: 1) Bark frequency warping, 2) equal-loudness pre-emphasis, and 3) intensity-to-loudness conversion. Perceptual log area ratios derived from PLP coefficients have linear spectral sensitivity and are more robust to quantization noise [29].

3.3 Cepstral Analysis

Cepstral analysis is a homomorphic transformation primarily used to convert a product expression into summation [3]. Under the assumption that short duration speech segments are quasi-stationary, speech signal is modeled as the output of slowly time-varying linear vocal tract filter excited by quasi-periodic pulses (during voiced speech) or random noise (during unvoiced speech) [3], [25]. This operation is linear convolution in time domain and multiplication in the frequency domain given by

$$\begin{aligned} s(n) &= e(n) * v(n) \\ S(\omega) &= E(\omega)V(\omega) \end{aligned} \tag{3.6}$$

where $s(n)$, $e(n)$ and $v(n)$ are the output speech samples, the excitation and the vocal tract impulse response, and $S(\omega)$, $E(\omega)$ and $V(\omega)$ are their spectra, respectively. The excitation and vocal tract characteristics can be deconvolved by taking the logarithm of the speech spectrum:

$$\log|S(\omega)| = \log|E(\omega)| + \log|\mathrm{V}(\omega)| \tag{3.7}$$

Since vocal tract characteristics (formant structures) change slowly compared to the excitation signal, the two spectra are quite different. Hence, they can be easily separated by linear filtering of the log speech spectrum, in which the low frequency components relate to the vocal tract envelope and the high frequency components to source information. The magnitude cepstrum is defined as the inverse Fourier transform of the logarithm of the magnitude spectrum given by [1], [3]

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|S(\omega)|\, e^{j\omega n} d\omega \tag{3.8}$$

where $c(n)$ is the n^{th} cepstral coefficient. Applications of cepstral analysis include pitch and formant estimation, and in speech coding (vocoders).

For speaker and speech recognition, cepstral coefficients derived from linear prediction coefficients or directly from the short-time speech spectrum are used. The derivation of LP cepstral coefficients (LPCC) is given by the recursive formula in (3.5). The popular FFT cepstral coefficients called Mel-frequency cepstral coefficients (MFCC) are computed using Mel-filter bank analysis. Filter bank analysis is employed because our hearing system is sensitive to a band of frequencies rather than to each single frequency [2], [26]. The Mel-scale filter bank is designed

to capture the perceptually significant spectral components of speech based on the human perception. In the linear frequency scale, the center frequencies of the filters are linearly spaced in the lower frequency and logarithmically spaced in the higher frequency. The Mel-scale frequency is related to the linear-scale frequency f in Hz by [3], [30]

$$\text{mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{3.9}$$

Figure 3.3 illustrates the computation of MFCC. First the input speech sample is segmented into frames of length N, and N-point FFT of each frame is computed. The magnitude of the spectrum is then weighted by Mel-filter bank. The output X_m of the m^{th} filter bank is given by:

$$X_m = \sum_{k=0}^{N-1} |S_n(k)| H_m(k), \quad m = 1, 2, \ldots, M \tag{3.10}$$

where $S_n(k)$ is the FFT of the speech frame, $H_m(k)$ is the m^{th} filter and M is the number of filters in the filter bank, typically 20. The MFCCs are calculated as discrete cosine transform (DCT) of the logarithm of the square of X_m, also called the log-energy, using the following formula [3]:

$$c_j = \sum_{m=1}^{M} \log\left(|X_m|^2\right) \cos\left[j\left(m - \frac{1}{2}\right)\frac{\pi}{M} \right], j = 1, 2, \ldots, M \tag{3.11}$$

where c_j is the j^{th} cepstral coefficient.

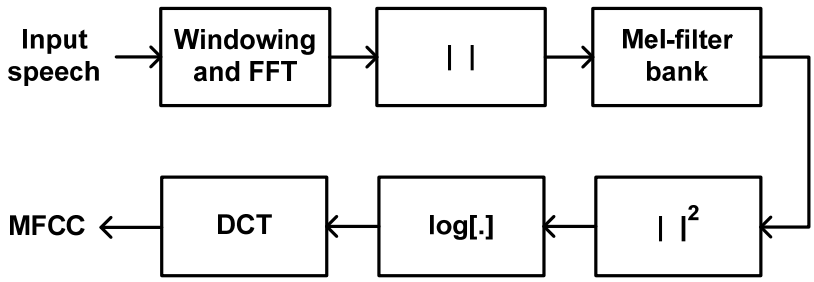

Fig. 3.3 MFCC extraction

LPCCs and MFCCs are both effective in representing speaker-dependent characteristics of speech. However, LPCCs are more often used because they are directly related to the formant structure of the vocal tract. Beside, cepstral analysis is more computation intensive as compared to LPC analysis. The advantage of cepstral analysis is that there is no assumed model of the vocal tract system and cepstral parameters can be modeled well by Gaussian mixture densities [2]. The first and second time derivatives of MFCC, called delta and delta-delta coefficients, respectively are also independent of the MFCCs and carry dynamic speaker information [2].

3.4 Harmonic Plus Noise Model (HNM) Parameters

HNM is a parametric model often used for synthesizing high quality, speech. In this model, the speech signal is assumed to be composed of a time-varying harmonic part and a modulated noise part [19], [31]. The synthesized speech signal $\hat{s}(t)$ is given as the sum of the two components:

$$\hat{s}(t) = s_{\mathrm{h}}(t) + n(t) \tag{3.12}$$

where $s_{\mathrm{h}}(t)$is the harmonic part and $n(t)$ is the noise part. The harmonic part accounts for the voiced components of speech and is represented as the sum of harmonically related sinusoids, given by

$$s_{\mathrm{h}}(t) = \mathrm{Re}\left\{ \sum_{k=1}^{K(t)} a_k(t) \exp\left(j\left[\int_0^t k\omega_0(\sigma)d\sigma + \theta_k \right] \right) \right\} \tag{3.13}$$

where $\omega_0(t)$, $a_k(t)$ and $\theta_k(t)$ are the time varying fundamental frequency, amplitude and phase of the k^{th} harmonic, respectively, and $K(t)$ is the number of harmonics required to represent the harmonic component. The noise part which represents the remaining unvoiced speech components is given by autoregressive model:

$$n(t) = w(t)\left[h(\tau, t) * g(t)\right] \tag{3.14}$$

where $g(t)$ is white Gaussian noise, $h(\tau, t)$ is a time-varying normalized all-pole filter and $w(t)$ is an energy envelope function.

HNM assumes that the lower band of speech spectrum can be represented solely by harmonics and the upper band solely by noise [31]. The bands are separated by a time-varying cut-off frequency called maximum voiced frequency. This provides a simple model for speech synthesis and modification. The complete HNM model requires the estimation of the harmonic part parameters, the maximum voiced frequency and the pitch from the speech samples. The noise part is obtained by subtracting the harmonic part from the speech signal.

The use of HNM parameters for speaker recognition was investigated by Gangula *et al* [7], [19]. Their work included the study of the maximum voiced frequency F_m, pitch F_0, the ratio of the noise part energy to the total energy A_n^2 / A^2, jitter in F_m (ΔF_m), jitter in F_0 (ΔF_0), and shimmer in intensity (ΔA_0). The motivation was that HNM parameters contain speaker-dependent information related to the vocal tract and vocal folds, and are less susceptible to noise. The variation of the HNM parameters within and across speakers was studied using histogram analysis, statistical moments, correlations and F-ratio tests. The investigation was carried out using a speech database of 8 sentences recorded from 10 speakers. The duration of the senetcnes varied from 2-3 s, and each sentence was spoken 5 times by each speaker. The results showed that

HNM parameters vary more across speakers than within each speaker, and they could be useful for speaker recognition.

Speaker recognition experiments were conducted using VQ modeling of the HNM and MFCC parameters extracted from short test utterances. The performance of the HNM parameters was comparable to that of MFCC features. Text-dependent tests using only HNM parameters resulted in recognition rates of 72%-100% for different test utterances. Recognition rates of 56%-78% were obtained using different training and testing utterances. Recognition tests using combination of MFCC and HNM parameters resulted in an improvement of up to 40% over MFCC features alone.

3.5 High Level Features

Spectral features of speech extracted from short segments are commonly used for speech and speaker recognition because they i) are easy to extract, ii) contain inherent speaker characteristic which are difficult to mimic, and iii) give high recognition accuracy with small to moderate amount of training data [2], [5], [6], [15], [32]. However, they are affected by noise and channel mismatch. Recent investigations using features extracted from learned speaking habits of speakers have shown promising results and preferred applications, such as the use of conversational patterns for detecting change of speaker in continuous speech, and for grouping speech segments of the same speaker [5], [6], [32].

High level features are those features beyond the acoustic level that capture linguistic and long-term stylistic behavior of the speaker such as prosody, pronunciation, phonetics, lexicon, syntax and semantics [5], [8], [33], [34]. These features are influenced by the socioeconomic and educational status, and childhood environment of the speaker. High level recognition systems extract a sequence of symbols such as pitch and timing patterns, n-grams of pronunciations, phonemes, and words, and use the frequency and co-occurrence of the these symbols to characterize speakers [5], [34]. High level features alone have not performed better than spectral parameters so far; however, they significantly improve recognition performance when they are used in combination with LPCC or MFCC features [5], [6]. The main disadvantages with the use of high level features are that they:1) are difficult to extract and often require automatic speech recognizers; 2) require a large amount of training data; and 3) are susceptible to mimics.

Chapter 4

STUDY OF STOP CLOSURE AND BURST DURATIONS

4.1 Introduction

For speaker recognition, spectral features such MFCCs and LPCCs perform well under suitable conditions. However, their performance gets degraded under noise and mismatched acoustic conditions. Besides, these features are extracted by assuming some simplified model of the speech production system. Both the all-pole model of the vocal tract and source-filter model of the overall speech production are approximations, and this may have a significant effect on the system performance. Recently, high level features such as word usage, prosody, pronunciation and duration have been used in combination with the short-time spectral features to improve recognition accuracy [15], [33], [34]. Ferrer *et al* [35] studied the use of word and phone duration features for speaker recognition using Gaussian mixture modeling, and reported significant improvement by combining with MFCC features. Shriberg *et al* [34] also used N-grams of syllable-level pitch, duration and energy features using support vector machine.

In this work, closure and burst durations in stops are investigated for speaker recognition using Gaussian mixture modeling. Since the closure and burst events are abrupt, speaker-dependent characteristics derived from stop phonemes may be difficult to be mimicked by impostors. Moreover, compared to other types of phonemes, stop closure and burst durations are less influenced by changes in speaking manner of speakers such reading, conversation, and stress [3], hence they could provide robust performance by combining with .MFCC features. In this chapter, the variation of stop closure and burst durations across speakers is studied using analysis of variance. Details of the analyses and results are presented.

4.2 Investigations

Stops are acoustically transient phonemes produced by a process of articulation involving complete closure and release of the vocal tract constriction, and their spectra generally have closure and frication burst regions. The closure interval is silence for most stops, but in some cases it can have a low frequency voice bar which results from the radiation of periodic glottal pulses through the vocal tract walls [3]. Manual labeling of the closure and burst durations for different stops in VCV utterances, spoken by a male. speaker MS8, is shown in Fig 4.1.The closures segments have similar durations across the different stops, while burst durations for

voiced stops (/b/, /d/, /g/) are very short as compared to that of unvoiced stops (/p/, /t/, /k/).The main reason for this is that it is difficult to visually separate the closure release burst from the unvoiced aspiration in case of unvoiced stops, and hence the burst duration is generally the same as the voice onset time.

Stop closure and burst durations are expected to vary among speakers due to anatomical and speaking style differences. Besides, stop closure and burst durations are affected by contexts (type of utterances), and speaking rate variations. Figure 4.2 shows closure and burst durations in the utterance /uku/ for different speakers. Figure 4.3 shows closure and burs durations for the utterances /aka/, /iki/ and /uku/, spoken by MS8. The plots show significant differences in the lengths of closure and burst durations across the speakers, vowels and stops. For these duration features to be used for speaker recognition, their variation among speakers should be greater than the variation within each speaker due to other factors. The effects of the various factors on stop closure and burst durations are studied using Fisher's F-ratio analysis of variances (ANOVA). Compared to bursts, closure segments have longer duration for most stops and they can have different ranges for different speakers. Besides, closure durations are easier to extract, hence they may be useful for speaker recognition and other speech processing applications.

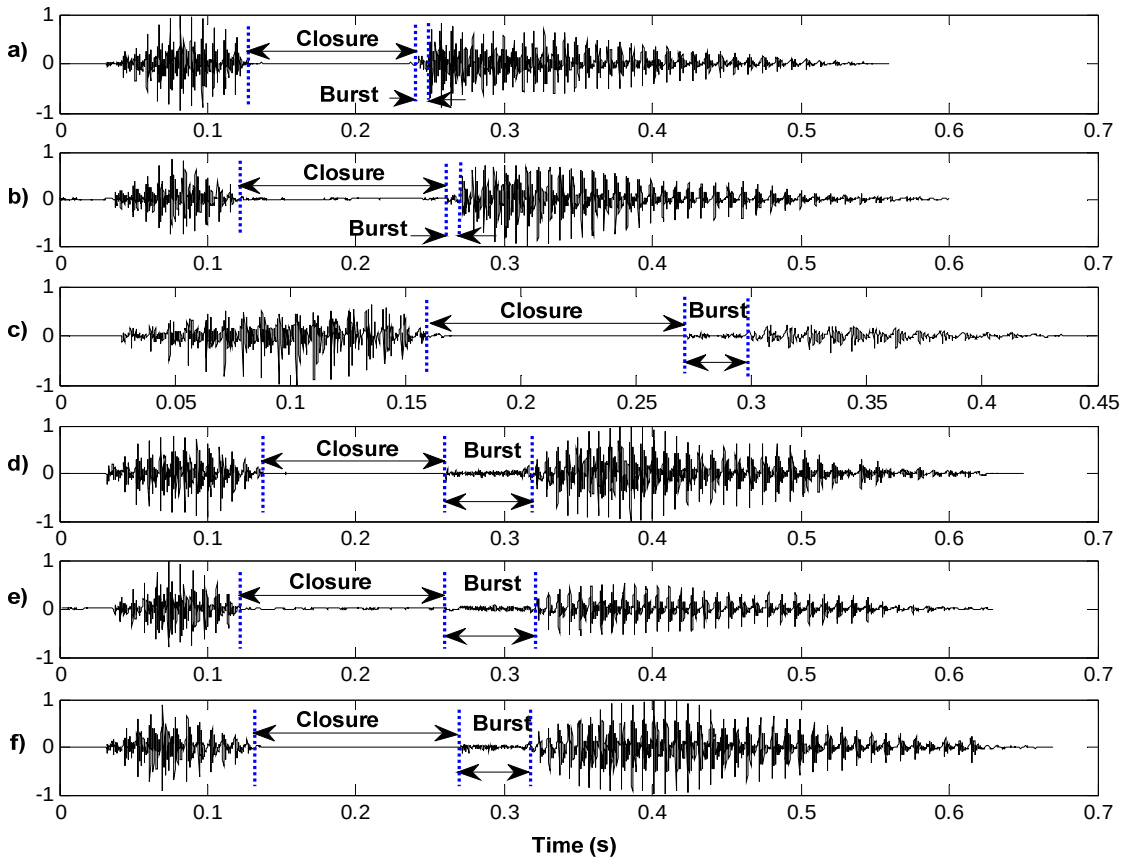

Fig. 4.1 Closure and burst durations for different VCV utterances spoken by male speaker MS8: a) /aba/, b) /ada/, c) /aga/, d) /aka/, e) /apa/, f) /ata/

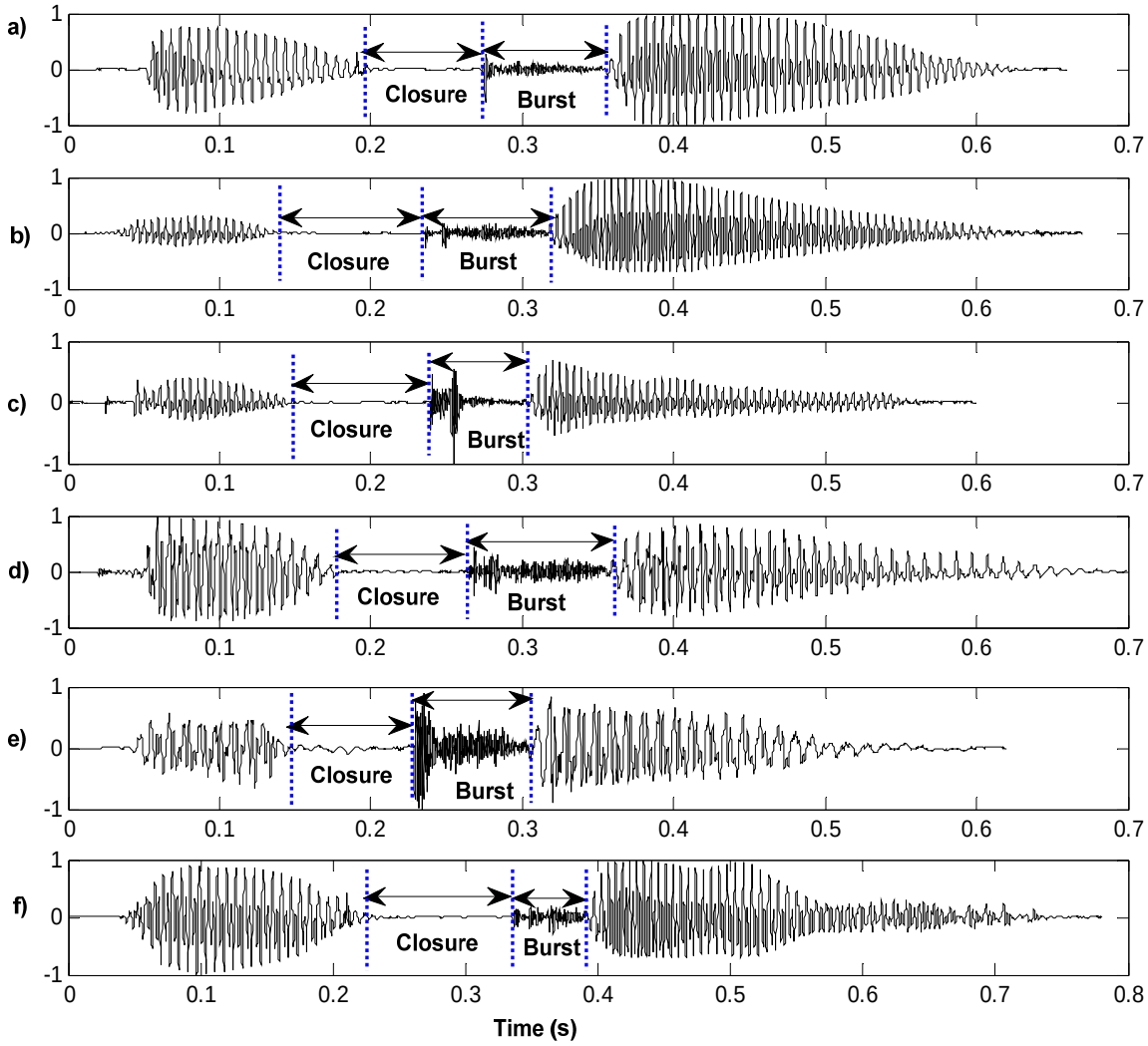

Fig. 4.2 Closure and burst durations in /uku/ spoken by different speakers: a) FS6, b) FS13, c) FS18, d) MS2, e) MS4, f) MS8; FS = female speaker, MS = male speaker

Analysis of variance is used for analyzing variations among sample groups due to a given set of independent factors, by comparing the estimates of the within and among group variances [36]. When the experiment involves only one independent variable, one-way ANOVA is used for testing if the group means are equal. The test is carried out using F-ratio, which is calculated as the ratio of the variance of group means to the average of within group variances [1], [3], [36]. A large F-ratio value indicates significant variation of the group means, compared to the intra-group variation; hence each group can be represented by a separate set of values of the parameter or vectors of its statistics (mean, variance). When there are more than one independent sources of variation (factors), factorial analysis of variance is used to examine the effects of the individual factors and their interactions.

The variation of stop closure duration across speakers was studied using one-way ANOVA at normal speaking rate. The analysis was carried out on recorded VCVs and sentence utterances, and TIMIT sentences. The vacation of burst duration across speakers was also studied using one-way ANOVA on TIMIT sentences. The effect of speaking rate variation was investigated using two-way ANOVA of closure duration at different speaking rates. This was conducted on sentences recorded at slow, normal, and fast speech rates. In addition, the effect of context variation on both closure and burst durations was studied on VCV utterances, using two-way ANOVA for different vowels and stops. In all cases, stop closure and burst durations were obtained manually, and the analyses involved computation and comparison of means, variances and F-ratio statistics for each factor. The analysis of variance method is discussed in the following section.

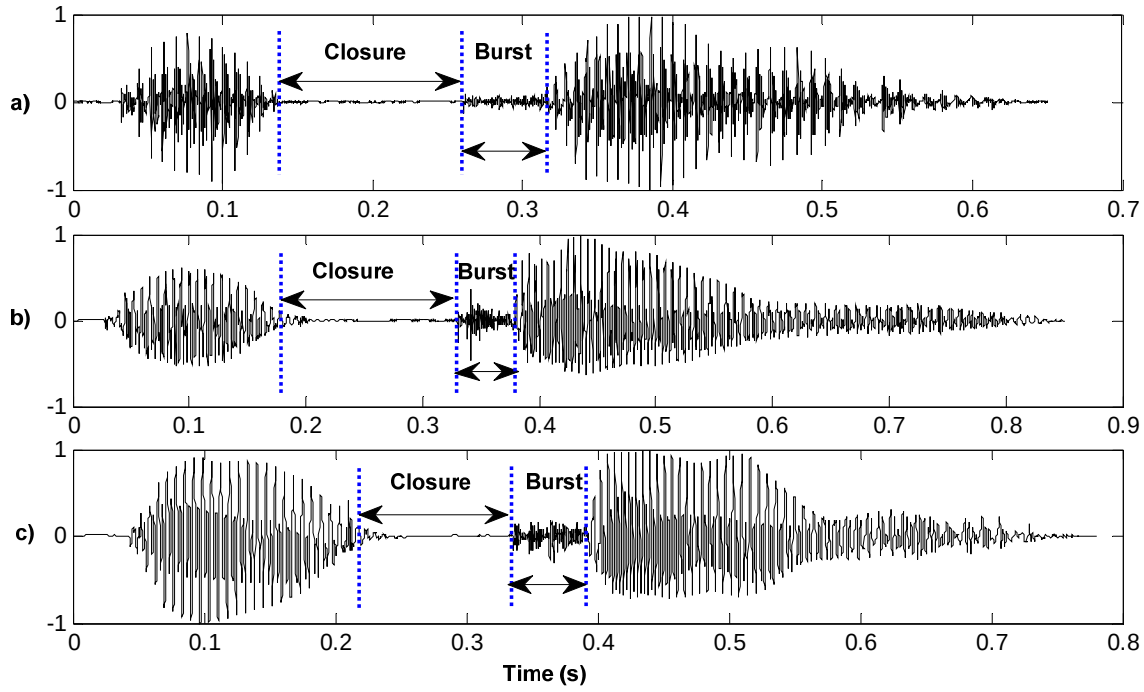

Fig. 4.3 Closure and burst durations for VCV utterances formed using /k/ with different vowels, spoken by male speaker MS8: a) /aka/, b) /iki/, c) /uku/

4.3 Analysis of Variance

One-way ANOVA involves estimation of the within and between group means and variances from the sample data. Given repeated observations of a specific stop closure duration in an utterance for each speaker in a test set, let $x_{s,t}$ represent the t^{th} trial duration of the stop closure for a particular speaker, s. The sample mean duration $\bar{x}_s$ of the stop closure for each speaker are given, respectively, by

$$\bar{x}_s = \frac{1}{N_r}\sum_{t=1}^{N_r} x_{s,t} \tag{4.1}$$

where N_r is the number of repetitions. The variance of the total observation data due to variations within each group, is measured in terms of the sum of squares within group (SS_W, computed as

$$SS_W = \sum_{s=1}^{S}\sum_{t=1}^{N_r}(x_{s,t} - \bar{x}_s)^2 \tag{4.2}$$

where S is the number of speakers. This value is also referred to as the error sum of squares (SS_e). The mean sum of squares within group (MSS_W) is given by [36]

$$MSS_W = \frac{SS_W}{N_T - S} \tag{4.3}$$

where $N_T = \sum_{s=1}^{S} N_r$ and N_T-S is the within group degree of freedom. The mean of closure duration means across groups (speakers) is calculated as

$$\bar{\bar{x}} = \frac{1}{S}\sum_{s=1}^{S}\bar{x}_s \tag{4.4}$$

The sum of the squares of group mean deviations from the inter-group mean is known as the sum of squares between groups (SS_B) and is given by

$$SS_B = \sum_{s=1}^{S} T_s\left(\bar{x}_s - \bar{\bar{x}}\right)^2 \tag{4.5}$$

SS_B is used as measure of the differences between groups. The mean of squares between groups (MSS_B) is obtained as

$$MSS_B = \frac{SS_B}{S-1} \tag{4.6}$$

where S-1 is the degree of freedom between groups. Finally, the F-ratio is calculated as the ratio of mean sum of squares between groups to means of sum of squares within groups, given by

$$F = \frac{MSS_B}{MSS_W} \tag{4.7}$$

Two-way analysis of variance is used to examine the individual and combined effects of two independent factors. Let the two factors be factor A with N_a levels, and factor B with N_b levels. If we assume the factors to be speakers and speaking rates, N_a will be the number of speakers and N_b will be the number of speaking rates. Suppose each speaker repeats a given utterance N_r times, i.e., number of replicates, at a given speaking rate. Let x_{ijk} be the k^{th} trial closure duration of a specific stop for speaker *i,* at speaking rate *j*. The total sum of squares

(SS_{Total}) is partitioned into sum of squares due to factor A (SS_a), factor B (SS_b), between groups (SS_{ab}), and within group (error) (SS_e). These values are calculated as follows [37]:

$$SS_{Total} = \sum_{k=1}^{N_r}\sum_{j=1}^{N_b}\sum_{i=1}^{N_a} x_{ijk}^2 - \frac{\left(\sum_{k=1}^{N_r}\sum_{j=1}^{N_b}\sum_{i=1}^{N_a} x_{ijk}\right)^2}{N_T} \tag{4.8}$$

$$SS_a = \frac{\sum_{i=1}^{N_a}\left(\sum_{k=1}^{N_r}\sum_{j=1}^{N_b} x_{ijk}\right)^2}{N_r N_b} - \frac{\left(\sum_{k=1}^{Nr}\sum_{j=1}^{Nb}\sum_{i=1}^{Na} x_{ijk}\right)^2}{N_T} \tag{4.9}$$

$$SS_b = \frac{\sum_{j=1}^{N_b}\left(\sum_{k=1}^{N_r}\sum_{i=1}^{N_a} x_{ijk}\right)^2}{N_r N_a} - \frac{\left(\sum_{k=1}^{Nr}\sum_{j=1}^{Nb}\sum_{i=1}^{Na} x_{ijk}\right)^2}{N_T} \tag{4.10}$$

$$SS_{ab} = \frac{\sum_{j=1}^{N_b}\sum_{i=1}^{N_a}\left(\sum_{k=1}^{N_r} x_{ijk}\right)^2}{N_r} - \frac{\left(\sum_{k=1}^{Nr}\sum_{j=1}^{Nb}\sum_{i=1}^{Na} x_{ijk}\right)^2}{N_T} - SS_a - SS_b \tag{4.11}$$

$$SS_e = SS_{Total} - SS_a - SSb - SS_{ab} \tag{4.12}$$

where $N_T = N_a N_b N_r$ is the total number of observations. The formulas for computing the corresponding mean values and two-way ANOVA statistics are summarized in Table 4.1.

Table 4.1 Computation of two-way ANOVA

Source	Sum of squares	Degree of freedom	Mean sum of squares	F-ratio
A (speakers)	SS_a	N_a-1	$MSS_a = SS_a / (N_a\text{-}1)$	MSS_a / MSS_e
B (speaking rates)	SS_b	N_b-1	$MSS_b = SS_b / (N_b\text{-}1)$	MSS_b / MSS_e
AB (interaction)	SS_{ab}	$(N_a\text{-}1)(N_b\text{-}1)$	$MSS_{ab} = SS_{ab} / (N_a\text{-}1)(N_b\text{-}1)$	MSS_{ab} / MSS_e
Error (within groups)	SS_e	$N_T - N_a N_b$	$MSS_e = SS_e / (N_T - N_a N_b)$	-
Total	SS_{Total}	N_T -1	-	-

The significance of the variation among groups is measured by F-distribution probability (P-value), computed using the degree of freedom between groups and degree of freedom within groups as the numerator and denominator degrees of freedom respectively [36]. This probability is calculated assuming all the group means are equal and it measures the probability of large random error. The significance of the variation is determined by comparing the P-value with a desired significance (risk) level, called α-value. A significance level of 5% (α=0.05) is often used. The variation between groups is statistically significant if the P-value is less than the α-

value [36]. Alternatively, the significance can be determined by comparing the F-ratio value against a critical F-value. The critical F-value for a given significance level is obtained from F-distribution tables using the numerator and denominator degrees of freedom. In this case, F-ratio value greater than the critical value shows significant variation between groups.

4.4 Results and Discussion

The ANOVA results obtained for closure and burst durations, using speakers, speaking rates, and vowel and stop types as independent factors, are given in the following sections.

4.4.1 One-Way ANOVA of Closure Duration

One-way ANOVA was used to analyze the variation of stop closure durations among speakers at normal speaking rate. A recorded database of 10 VCV utterances and 6 sentences from 5 male speakers was used for this purpose. The data base contains 7 trials of each utterance from each speaker and was collected over 2 days using the same microphone. A list of the utterances is given in Table 4.2. Closure durations were obtained manually from waveform plots and spectrograms obtained using "PRAAT". Only stop closures which are relatively easy to measure were taken to reduce measurement errors. Within and between speaker means and variances, and F-ratio values of closure durations were calculated for each stop in each utterance (text-dependent). Text-independent F-statistics were also obtained for the recorded VCVs and sentences separately.

able 4.2 List of utterances used for one-way ANOVA of closure duration

Label	Utterance
VCV01	/aba/, /ada/, /aka/ /apa/, /ata/,
VCV02	/ibi/, /idi/, /iki/, /ipi/, /iti/
S01	The editor dropped the paper.
S02	The capital city is highly populated.
S03	She picked up the package.
S04	She abandoned the baby.
S05	He photocopied the book.
S06	We abided the educated

Table 4.3 Means and square root of variances of text-independent stop closure durations in VCV utterances and sentences (Sp = speaker, var = variance)

Utterance	Mean (in ms)					$\sqrt{\text{var}}$ (in ms)				
	Sp1	Sp2	Sp3	Sp4	Sp5	Sp1	Sp2	Sp3	Sp4	Sp5
VCVs	113.	100.5	105.5	85.8	92.8	62.8	7.9	15.4	14.1	13.4
Sentence	70.3	56.7	60.2	51.2	57.6	21.5	24.7	18.3	12.6	17.4

Table 4.4 ANOVA of text-independent stop closure durations in VCV utterances and sentences. ($\alpha = 0.05$, SS_W = sum of squares within speakers, SS_B = sum of squares between speakers, MSS_W = mean of squares within speakers, MSS_B = mean of squares between speakers)

Utterance	SS_W (ms)	MSS_W (ms)	SS_B (ms)	MSS_B (ms)	F-ratio	P-value	$F_{critical}$
VCVs	53382.5	154.7	32617.2	8154.3	52.7	1.3e-34	2.4
Sentences	345746.0	375.4	37321.4	9330.4	24.9	1.5e-19	2.4

Table 4.3 shows text-independent mean and square root of variance of closure durations for the 5 speakers. The corresponding ANOVA results are given in Table 4.4. Text-dependent means and square root of variances, and ANOVA results are summarized in Table A.1 and Table A.2 in the appendix. For a specific stop closure in a given utterance, the mean closure durations for different speakers were quite different; the mean sum of squares within each speaker were also smaller than the between speaker mean sum of squares. The F-ratio values were significantly greater than the critical F-values and the P-values were much smaller than the 5% significance level. This indicates that there is significant variation of stop closure duration among different speakers, and these variations are larger than the intra-speaker variations. Therefore, closure duration is speaker-dependent, and hence a potential candidate for text-dependent or text-prompted speaker-recognition. There is also a large variation in the mean closure durations for a specific stop in different utterances of the same speaker which shows high context dependency of the parameter.

Text-independent ANOVA of stop closure and burst durations were carried out for 7 speakers using 7 sentences from TIMIT database. Each sentence was spoken once by each speaker. Closure and burst duration were obtained from the manual transcriptions. The mean and square root of variance values and summary of the ANOVA results are given in Table 4.5 and Table 4.6, respectively. The closure and burst duration means are not significantly different among the speakers, and the variances for each speaker are large. On the other hand, the F-ratio values are considerably greater than the critical value, which indicates that there is an overall variation of stop closure and burst durations across speakers. Figure 4.3 and shows the normalized histogram and 3-component Gaussian mixture approximation of stop closure durations for the 7 speakers. The histogram of each speaker has several peaks of different heights. This indicates that the closure duration values for each speaker may be concentrated on certain regions of the time axis, which correspond to the different stops. There is a difference in the location and height of peaks, and the overall pattern of the histogram and GMM contours among the speakers. The results show that distribution of stop closure durations varies across speakers and it may be used for text-independent speaker recognition. Figure 4.4 shows the distribution of stop burst durations

across the speakers. In this case, the contours for the different speakers are closely spaced, which indicates that burst durations may not be useful for text-independent speaker recognition.

Table 4.5 Means and square root of variances of stop closure and burst durations in TIMIT sentences

Parameter		Speakers						
		FAKS0	FHEW0	MJFC0	MKLT0	MNLS0	MRPC0	MTMR0
Closure duration	Mean (ms)	66.4	65.0	50.5	41.2	48.8	53.0	63.9
	$\sqrt{\text{var}}$ (ms)	27.8	21.0	21.8	16.3	19.5	19.8	22.9
Burst Duration	Mean (ms)	25.4	28.0	32.5	38.8	30.5	40.7	29.2
	$\sqrt{\text{var}}$ (ms)	15.6	18.2	19.2	19.3	17.2	23.9	14.1

Table 4.6 ANOVA of stop closure and burst durations in TIMIT sentences, $\alpha = 0.05$

Parameter	Source of Variation	Sum of squares (ms)	Degree of freedom	Mean of squares (ms)	F-ratio	P-value	F-critical
Closure Duration	Among speakers	31872.2	6	5312.0	11.4	8.9e-12	2.1
	Within speakers	184438.6	396	465.8	-	-	-
Burst duration	Among speakers	9386.7	6	1564.4	4.58	1.8e-04	2.13
	Within speakers	108589.2	318	341.5	-	-	-

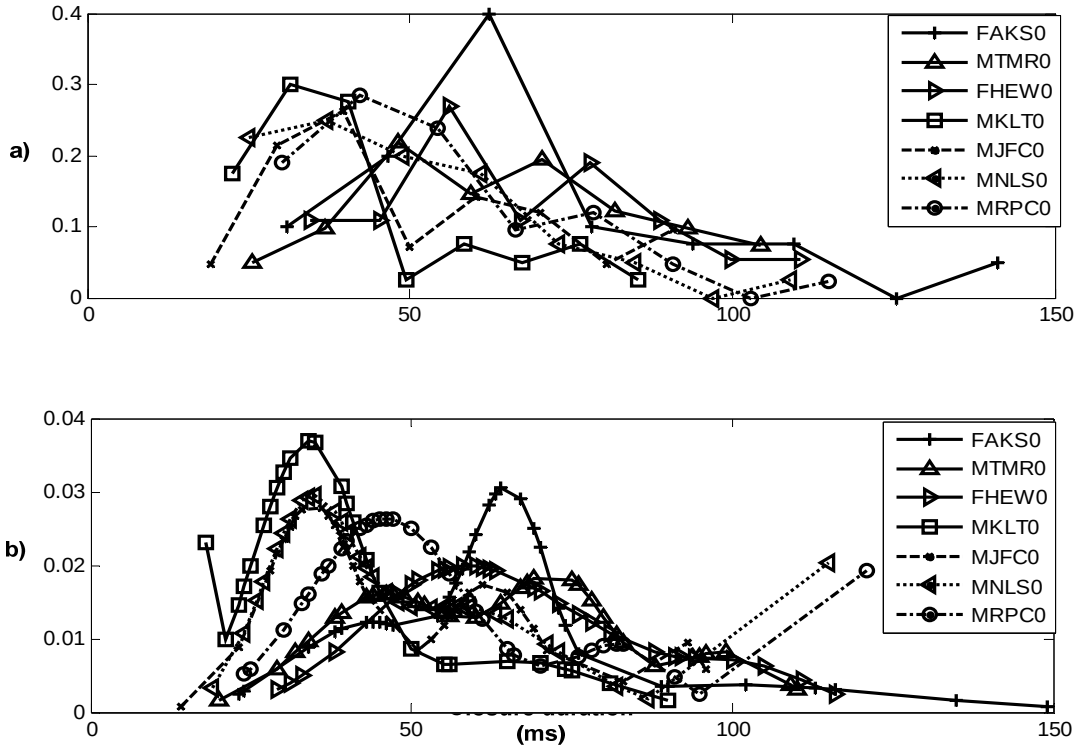

Fig. 4.4 Distribution of stop closure duration for 7 speakers from TIMIT database: a) normalized histogram, b) approximation using 3-compnent GMM

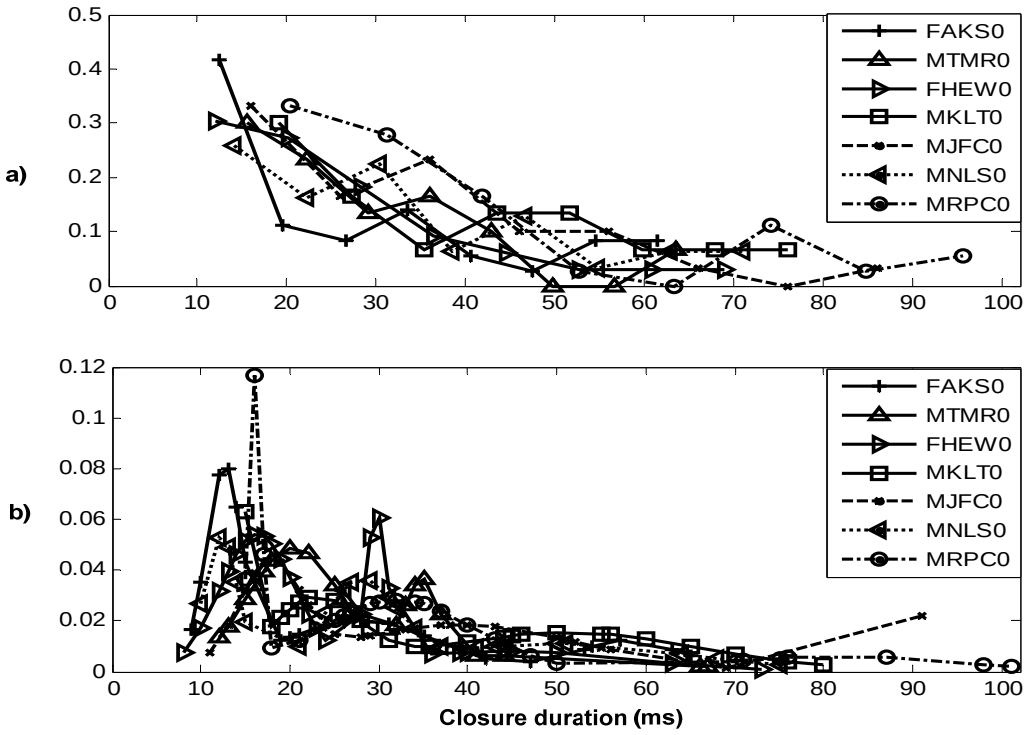

Fig. 4.5 Distribution of stop burst duration for 7 speakers from TIMIT database: a) normalized histogram, b) approximation using 3-compnent GMM

4.4.2 Two-Way ANOVA of Closure Duration at Different Speaking Rates

The variation of stop closure duration with speakers and speaking rates was studied using two-way ANOVA at different speaking rates. For this purpose, four sentence utterances consisting of stops were recorded at slow, normal and fast speaking rates from 3 male speakers. Each sentence was spoken five times by each speaker for each speaking rate. Stop closure durations were obtained manually from the waveforms and spectrograms using "PRAAT". A list of the sentences is given in Table 4.7. Two-way ANOVA of closure durations was computed in both text-dependent and text-independent modes, with speakers and speaking as the independent sources of variation. The text-independent means and square root of variances, and ANOVA summary are given in Table 4.8 and Table 4.9 respectively. The results for the individual utterances are summarized in Table A.3 and Table A.4.

Table 4.7 List of utterances used for two-way ANOVA of closure duration at different speaking rates

Label	Utterance
S11	The editor photocopied the book.
S12	The capital city is highly populated.
S13	She abandoned the abacus.
S14	She picked up the package

Table 4.8 Means and square root of variances of stop closure durations across speakers and speaking rates in sentences" (Sp = speaker, var = varaince)

Speaking Rate	Mean of speakers (ms)				$\sqrt{var}$ of speakers (ms)			
	Sp1	Sp2	Sp3	Total	Sp1	Sp2	SP3	Overall
Slow	97.2	147.7	91.0	112.0	43.2	52.7	35.6	51.0
Normal	59.4	69.0	62.8	63.7	18.1	18.6	22.6	20.2
Fast	44.9	50.8	51.0	48.9	16.9	14.4	18.6	16.9
Overall	67.2	89.2	68.2	74.9	36.2	53.6	31.4	42.7

Table 4.9 Two-way ANOVA of stop closure durations across speakers and speaking rates in senetences ($\alpha = 0.05$, Sp = speaker)

Source of variation	Sum of squares (ms)	Degree of freedom	Mean of squares (ms)	F-ratio	P-value	$F_{critical}$
Rate	522207.8	2	261103.9	296.2	2.6e-94	3.0
Speakers	73781.5	2	36890.8	41.9	6.5e-18	3.0
Interaction	86550.8	4	21637.7	24.5	4.7e-19	2.4
Within	626659.0	711	881.4	-	-	-

Observing at the table of means and square root of variances, we see that there is significant variation in mean closure durations due to both speaker and speaking rate variation. The F-ratio values for speaking rate, speaker and their interaction are all larger than the critical F-values in both text-dependent and independent modes. However, for most of the stop closures, the differences in mean closure durations of the same speaker for different speaking rates are larger than the differences due to speakers for a given speaking rate. The speaking rate F-ratios are also larger than the speaker F-ratios and the interaction F-ratios are mostly lowest. This means that the variation of stop closure duration due to speaking rate is higher than the variation among speaker, and that the performance of a speaker recognition system based on duration features might be poor under speaking rate changes. This problem can be reduced by normalizing the duration values by an estimate of the speaking rate. On the other hand, speaking rate is speaker-dependent [2] and it could be one reason for the variation of duration features among speakers, hence normalization could remove the speaker-dependent information.

4.4.3 Two-Way ANOVA of Closure and Burst Durations in VCV Utterances

In the previous sections, it was found that text-dependent stop closure duration significantly varies across speakers and could be helpful for text-dependent speaker recognition. The text-independent results on both the recorded and TIMIT data also showed considerable difference between speakers, which motivates us to investigate the stop duration parameters for text-independent speaker recognition. In order to study the variation of duration features with different contexts and

Table 4.10 Two-way ANOVA of stop closure and burst durations for different vowels (5 female and 5 male speakers, $\alpha = 0.05$)

Parameter	Source of variation	Sum of squares (ms)	Degree of freedom	Mean of squares (ms)	F-ratio	P-value	$F_{critical}$
Closure duration	Vowels	4238.1	2	2119.1	5.6	5.0e-03	3.1
	Speakers	72839.6	9	8093.3	21.3	7e-23	1.9
	Interaction	18521.9	18	1029.0	2.7	5.0e-04	1.7
	Within	56959.0	150	379.7	-	-	-
Burst duration	Vowels	2058.2	2	1029.1	1.0	0.4	3.1
	Speakers	12151.4	9	1350.2	1.3	0.3	1.9
	Interaction	4401.5	18	244.5	0.2	1.0	1.7
	Within	157859.0	150	1052.4	-	-	-

Table 4.11 Two-way ANOVA of stop closure and burst durations for different stops (5 female and 5 male speakers , $\alpha = 0.05$)

Parameter	Source of variation	Sum of squares (in ms)	Degree of freedom	Mean of squares (in ms)	F-ratio	P-value	$F_{critical}$
Closure duration	Stops	8680.7	5	1736.1	3.89	2.6e-03	2.29
	Speakers	72839.6	9	8093.3	18.14	1.0e-18	1.96
	Interaction	17495.6	45	388.8	0.87	0.7	1.47
	Within	53542.7	120	446.2	-	-	-
Burst	Stops	122900.0	5	24580.0	115.16	4.3e-44	2.29
	Speakers	12151.4	9	1350.2	6.33	2.7e-07	1.96
	Interaction	15806.0	45	351.2	1.65	1.7e-02	1.47
	Within	25612.7	120	213.4	-	-	-

across speakers, we performed two-way ANOVA of stop closure and burst durations for different stops and vowel contexts. A database of 18 VCV syllables spoken by 5 female and 5 male speakers was used. The syllables were formed from 3 vowels (/a/, /i/, /u/) and six stops (/b/, /d/, /g/, /k/, /p/, /t/) and each syllable was spoken once by every speaker at normal speaking rate. Closure and burst durations were obtained manually. Two types of two-way ANOVA were computed: (i) using vowel type and speakers as the independent variable, ignoring stop type, and (ii) using stop type and speakers as the independent variables, ignoring vowel type.

A summary of the two-way ANOVA of closure and burst durations for different vowels are given in Table 4.10. The means and square root of variances are given in Table A.5. The results show that there is large variation in mean closure duration due to both vowel type and speaker change. The F-ratio values are also greater than their respective critical values, which indicate that both factors have significant effect on closure duration. However, the F-ratio value due to speaker variation is larger than the F-ratio value due to vowel type, and the interaction F-ratio is relatively small. This shows that stop closure durations vary more significantly across speakers than across different vowels and the speaker-vowel dependence has relatively small

effect. In case of burst duration, there is small variation in the speaker means across different vowels; relatively, there is larger difference across speakers. However, the square root variances for each vowel and each speaker are quite large and the F-ratios are smaller than the critical F-values, which suggest that there is no remarkable difference in mean burst durations across speakers for different vowels.

Table 4.11 shows the results for the ANOVA of closure and burst durations for different stops. The mean and square root of variances are given in Table A.6. For most of the stops, there is a large difference in the mean closure durations of the same stop across the speakers. There is also a considerable variation in the mean closure durations of the same speaker for different stops. The F-ratio values for both speakers and stops are larger than the critical values, indicating dependence of the closure duration on both stop type and speakers. One important observation here is that the F-ratio due to speakers is much higher than the value due to stop type; hence stop closure duration can be used for stop-independent speaker recognition. Looking at the results for burst duration, we find large difference in mean duration across stops and relatively small across speakers, and the F-ratio due to stop type is significantly large compared to the F-ratio due to speaker variation. This suggests that the burst duration is highly dependent on the type of stop, and it could be important for stop identification; but its use for speaker recognition may be limited to text-dependent systems.

4.5 Summary

Stop closure and burst durations are generally affected by speakers, type of utterances (contexts) and speaking rate variations. The one-way analysis of variance results showed that stop closure and burst durations vary across the speakers and thus convey speaker-dependent information. Therefore, they can be used in combination with the short-time spectral features for improving speaker recognition. The two-way ANOVA of closure duration at different speaking rates indicated that closure duration is highly dependent on speaking rate and its performance may be degraded when there is a mismatch in speaking rate between the training and testing utterances. The ANOVA results using different vowels and stops showed that the variation of stop closure duration across speakers is large as compared to the variation due to vowel and stop, while burst duration depends mainly on the stop type. The results show that closure and burst durations are speaker-dependent, and compared to burst duration, closure duration conveys more speaker information. Both parameters are generally context-dependent, and they may be more important for text-dependent systems. However, a text-dependent system using duration features requires a speech recognizer for extracting the individual phone or word duration values, which complicates the system. In this project, speech landmark detection is used to extract text-independent duration features for speaker recognition.

Chapter 5

DETECTION OF SPECTRAL TRANSITIONS

5.1 Introduction

Recent studies have shown that speech landmarks contain significant perceptual and acoustic information that signify boundaries of major articulatory, acoustic, and phonetic features [38], [39]. These prominent spectral transitions have been used for knowledge-based automatic speech recognition using landmark detection, to segment the speech signal into linguistic units. In addition to linguistic information, speech signal contains features that characterize speakers and these features might be effectively extracted from selected portions of the speech signal. This work proposes the use of automatic landmark detection for extracting speaker-dependent temporal features in transient speech segments, and investigates the rate of spectral variation for locating the spectral transitions.

Speech landmarks are regions in speech signal where the acoustic-phonetic properties change abruptly. They include closure and release instants in stops, nasals and fricatives, and points of maximum intensity in vowels [38]. Short-time spectral and temporal features in the speech signal such as rate-of-change of energy at various frequency bands, formant transitions, voicing and duration parameters are used to detect speech landmarks. Several automatic landmark detection techniques for speech recognition have been reported [38] - [44]. A brief review of some of these methods is given in this section.

Liu [38] developed a method for detecting abrupt landmarks for distinctive feature-based speech recognition, based on the first difference of short-time dB energy in specific frequency bands. The spectrum of the input speech samples computed using 6 ms Hanning window every 1 ms, was divided into 6 frequency bands (0.0–0.4, 0.8–1.5, 1.2–2.0, 2.0–3.5, 3.5–5.0, 5.0–8.0 kHz) which correspond to spectral prominence of different phonemic classes. Peaks in the smoothed squared magnitude of the spectrum in each band were used to represent the band energy for tracking abrupt acoustic events using coarse and fine processing. In the coarse processing, energy waveform in each band was smoothed using 20 ms average window, and a rate o rise function was calculated as the first difference of the dB energy every 1 ms with 50 ms time step. The same operation was applied in the fine processing using 10 ms smoothing window and 26 ms time step. The energy rate-of-rise peaks from the two levels of processing were combined with duration, phonetic-class and articulatory constraints to detect three types of landmarks: glottal which

marked the start and end of voicing, burst which indicated the closure and release of stops and fricatives, and sonorants which located the closure and release of nasals and glides. The algorithm was tested on a speech database of 40 sentences constructed from 250 words (69% monosyllabic, 30% bisyllabic, 1% trisyllabic), recorded in a quiet room with signal-to-noise ratio of 30 dB. A landmark was considered correctly detected if it was within ±30 ms of manually labelled landmarks. The overall detection rates were 44%, 73%, 83%, and 88% within 5, 10, 20, and 30 ms of manually labelled landmarks respectively. Tested on TIMIT clean speech database, recognition accuracies of 98% and 90% were reported for glottal and burst landmarks, respectively, while sonorant landmarks were found to be difficult to detect.

Motivated by the fact that listeners use temporal cues to understand spectrally degraded speech, Salomon *et al*. [39] studied signal envelope and periodicity features for landmark detection using auditory filter bank analysis. The envelope of the signal from each channel was obtained using Hilbert transform to avoid excessive smoothing. Periodicity, pitch, and energy onset/offset measures were extracted for each channel. The periodicity measure included the classification of the signal envelope as silence, periodic, or aperiodic. Energy onset/offset measures were computed as the first difference of log energy between adjacent non-overlapping rectangular windows with adaptively varying window length. Measurements from each channel were combined to determine overall onset/offset, and periodic and aperiodic energy parameters. These parameters were used to detect three types of landmarks related to voicing onset/offset, sonorant consonants, and obstruent consonant boundaries. The method was evaluated using TIMIT speech database, and the overall detection rate was 80.2% with 8.7% insertion on clean speech, and 76% with 36.6% insertion on noisy speech.

Howitt [41] used low frequency energy measures (300 Hz – 900 Hz) to detect vowel landmarks. The input signal was sampled at 10 kHz and 256-point DFT was computed using 16 ms Hamming window every 5 ms. Energy measure was calculated as the sum of the log of the squared magnitude spectrum weighted by a trapezoidal window along the frequency axis. Peaks and dips in the energy function were found using a recursive convex hull peak picking algorithm [43]. Dips below a lower threshold were taken to be vowel boundaries and peaks above an upper threshold were considered as vowel landmarks (syllable centers). Testing on selected TIMIT database resulted in detection rate of 76% with 13.6% insertions. A landmark was considered correct if it was located within the manually transcribed vowel segment.

Jayan *et al*. [42] used modification (time-scaling) of vowel, consonant, and vowel-consonant (VC) transition segments for improving perception by hearing impaired listeners. They used rate-of-change (ROC) of peak energy and centroid frequency measures in a number of bands to detect VC transitions. The input speech was sampled at 10 kHz and 512-point DFT was computed using 6 ms Hanning window every 1 ms. The spectrum was split into five non-

overlapping frequency bands: 0–0.4, 0.4–1.2, 1.2–2.0, 2.0–3.5, 3.5–5.0 kHz. First differences of dB peak energy and centroid frequency were computed using 2 ms time step to obtain two ROC functions in each band. The absolute ROC values of peak energy and centriod frequency were multiplied and normalized in order to get a single ROC contour for each band. Finally, the product ROC values in the five bands were added and the resulting signal was averaged using 20 ms window to obtain an overall transition index for locating spectral transitions. Testing was done on 180 VCV utterances from 5 male and 5 female speakers, composed of 6 stops (/b/, /d/, /g/, /p/, /t/, /k) in 3 vowel contexts (/a/, /i/, /u/). The detection rates were 34%, 49%, 87%, 95%, 96%, and 100% within 3 ms, 5 ms, 10 ms, 15 ms, 20 ms and 30 ms respectively, of manually labelled spectral transitions.

In a recent work, Jayan and Pandey [46] used rate-of-change parameters obtained from Gaussian mixture modeling of short-time log magnitude spectrum for detecting stop landmarks at improved temporal accuracy. Fourier transform of the input speech, sampled at 10 kHz, was computed using 512-point DFT and 6 ms Hanning window with 1 ms shift. The magnitude spectrum was smoothed using a 50-point median filter along the frequency axis. Log of the smoothed magnitude spectrum was modelled as a weighted sum of Gaussian functions. The weights, means, and variances of the component functions were obtained using iterative expectation maximization (EM) likelihood estimation. Typically, the spectrum was modelled using four Gaussian functions with equal initial weights, and the means and variances were initialized to the average formant frequencies and extreme bandwidths of vowel. Contours of mean, variance and amplitude of the Gaussian envelope at mean location of each component was filtered using 30-point median filter to suppress transitions during steady state segments. First difference of each of these parameters was computed using 2 ms step for each Gaussian component. Rate-of-change (ROC) functions for mean, variance and envelope amplitude were computed as the sum over the four components of the absolute first difference values of the respective parameter at the same frame location. The product of the three ROC functions, in combination with spectral flatness measure and voicing onset/offset detector, was used for detecting stop landmarks. The method was tested on VCV utterances and TIMIT sentences. Stop burst landmarks in VCVs were detected at rates of 90%, 92%, 93%, 96% and 98% within 5, 10, 15, 20 and 30 ms of the manual landmarks, respectively. The detection rates on TIMIT sentences were 73%, 90%, 95%, 97% and 98% for burst landmarks, 19%, 40%, 63%, 80% and 90% for closure landmarks, and 45%, 82%, 91% and 96% for voicing onset landmarks.

5.2 Detection of Spectral Transitions from Magnitude Spectrum

Spectral transitions are instants in speech signal where the signal properties change abruptly and significantly in some region of the speech spectrum [38]. They are particularly manifested by

rapid changes in signal intensity or some patterns of it (such as periodicity and location of spectral prominence); hence they could be detected based on parameters extracted from the magnitude spectrum. Rate of change of short-time energy or some defined measure of signal level is often used for detecting instants of signal transitions. We have studied the rate-of-change (ROC) of mel-filtered magnitude spectrum for detecting the spectral transitions. Before selecting this parameter for our investigation, we conducted a preliminary assessment on the use of ROC of (i) squared magnitude spectrum, (ii) spectral energy in 5 bands, and (iii) mel-filtered squared magnitude spectrum. A single ROC of the magnitude spectrum is expected to detect spectral transitions with abrupt intensity change. However, it may fail to indicate transitions between sounds that have different spectral locations but may not result in significant magnitude variation. The use of multiple ROC parameters solves this problem by looking for landmarks in specific frequency bands. It also gives information about the type of landmark, which is useful for speech recognition. However, band specific ROC function can be sensitive to narrow band noise and harmonic transitions within voiced phonemes.

For the preliminary observation, the magnitude spectrum of the input signal, sampled at 10 kHz, was computed using 6 ms Hanning window and 512-point DFT every 3 ms. The short-time window suppresses harmonic details and provides fine time resolution as required for improved temporal accuracy in locating the spectral transitions. However, wideband spectrograms of voiced speech segments show vertical striations corresponding to individual pitch periods which might lead to the detection of false landmarks. In order to smooth these variations, the magnitude spectrum was filtered using median and mean filters along the time axis. Typically, 7-point median filter was found to give good spectral smoothing without affecting the temporal resolution. The dB first difference of the smoothed magnitude spectra were computed using 3 ms time step. A single ROC function was obtained as the mean of the first 25% largest absolute values in each first difference frame. Taking mean of the largest difference values, instead of the peak in each frame, reduces the possibility of false landmark detection due to narrow band noise. To obtain 5 band ROC contours, the spectrum was split into the following non-overlapping bands: 0-0.4, 0.4-1.2, 1.2-2.0, 2.0-3.5, 3.5-5 kHz, denoted as band1, band2, band3, band4 and band5 respectively. A single ROC was obtained for each band in the same way as used to get the ROC for the entire magnitude spectrum.

The techniques were tested on VCV utterances from two female speakers (FS6 and FS18) and one male speaker (MS2). Two-dimensional plots of the first difference and ROC contours were used for visual observation and comparison. Typical results for the sound |aba| spoken by female speaker FS6, obtained using 7-point median filter and a mel-filter bank with lowest filter bandwidth of 200 Hz, are given in Fig 5.1-Fig 5.3. The first difference of the raw magnitude spectrum resulted in high intensity difference values corresponding to variations between

successive pitch periods in voiced speech segments. Its ROC does not distinctly indicate the spectral transitions, especially the vowel ends. Median filtering reduces the periodic spectral variations and noise along the time axis giving relatively improved ROC peaks at spectral transitions. But the results are not satisfactory due to the spectral discontinuities along the frequency axis.

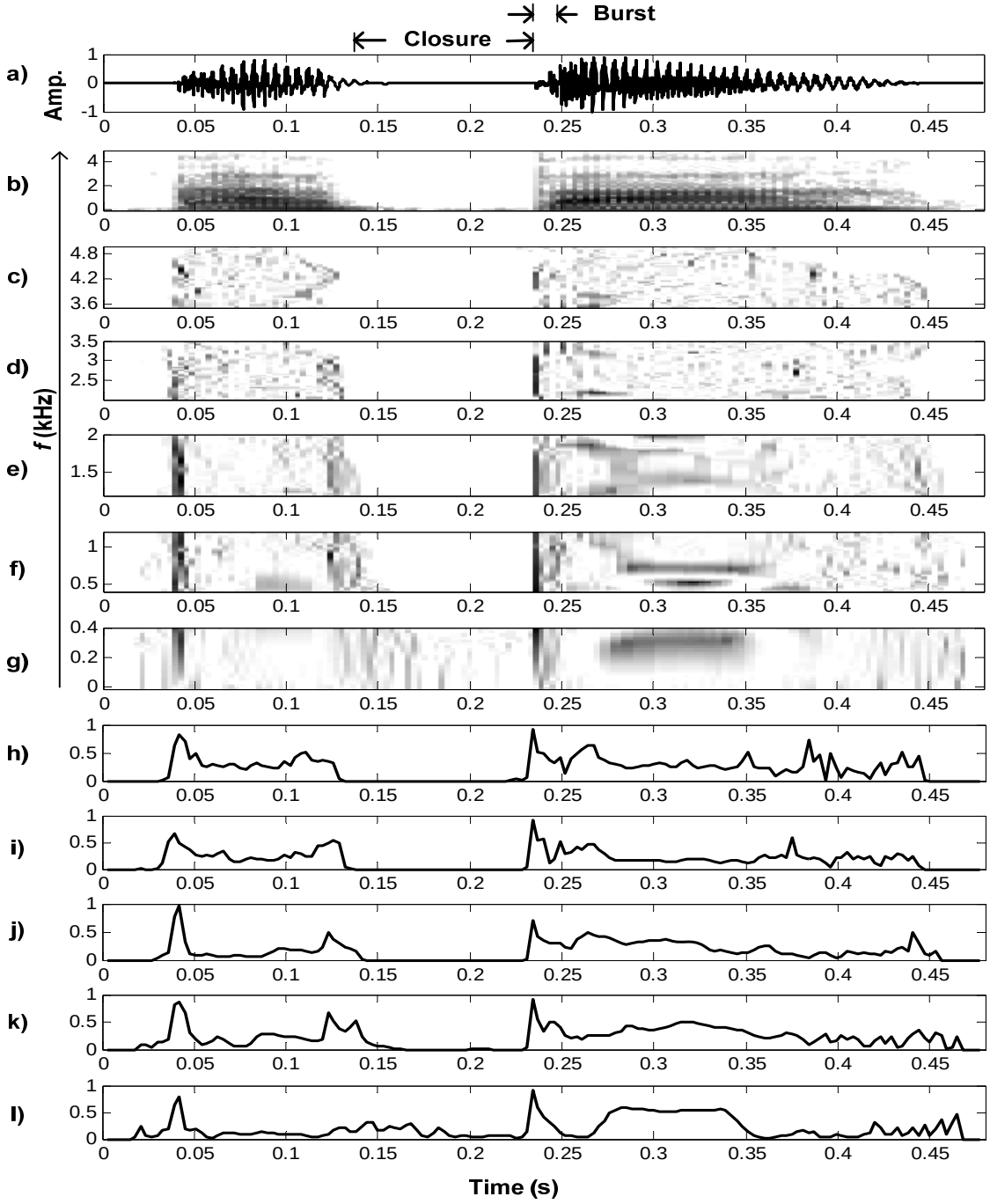

Fig. 5.1 Detection of spectral transitions for utterance /aba/ using magnitude spectrum in 5 bands: a) speech waveform, b) spectrogram, c) band5 first difference, d) band4 first differences, e) band3 first difference, f) band2 first difference, g) band1 first difference, h) band5 ROC, i) band4 ROC, i) band3 ROC, k)band2 ROC, l) band1 ROC

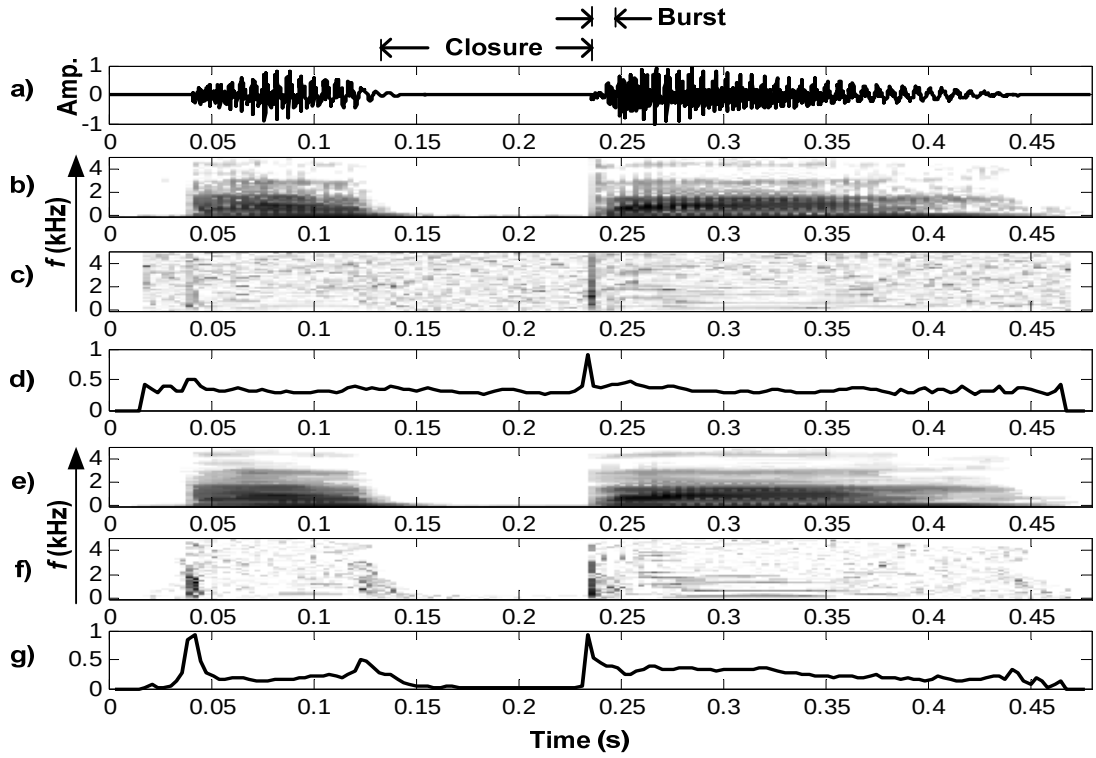

Fig. 5.2 Detection of spectral transitions for the utterance /aba/ using magnitude spectrum: a) speech waveform, b) spectrogram, c) first difference, d) ROC, e) median filtered spectrogram, f) first difference, g) ROC

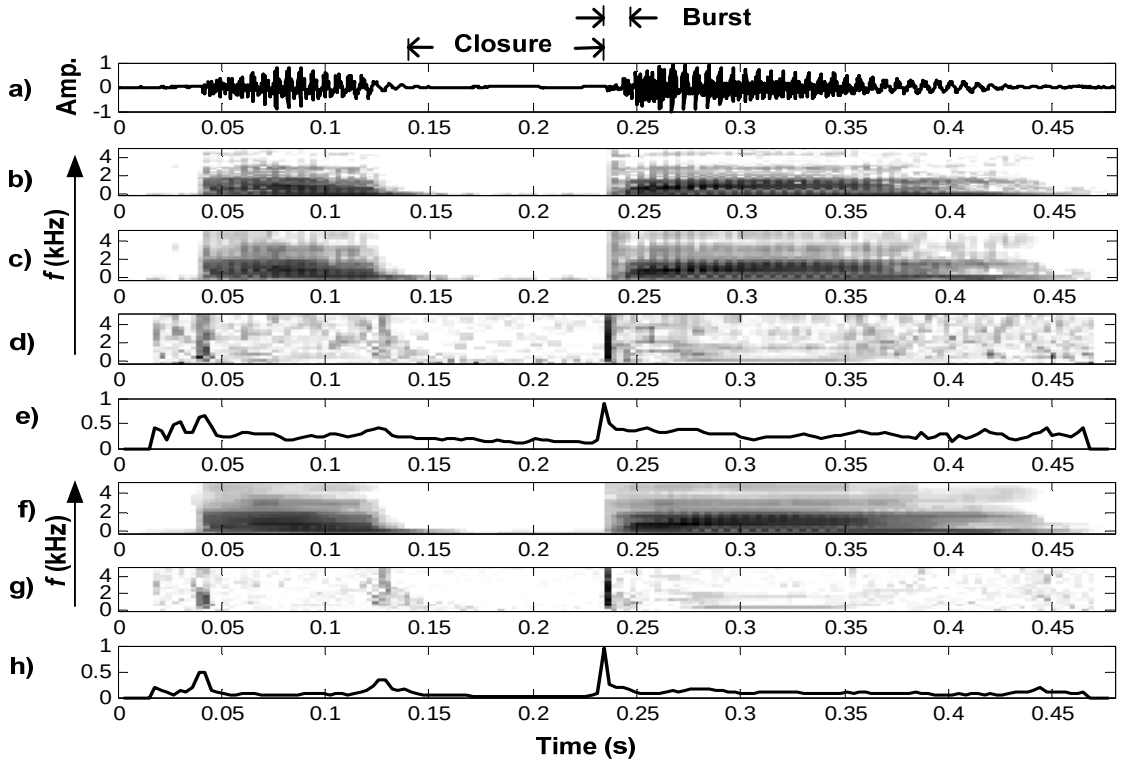

Fig. 5.3 Detection of spectral transitions for the utterance /aba/ using mel-filtered magnitude spectrum: a) speech waveform, b) spectrogram, c) mel-filtered spectrogram, d) first difference, e) ROC, f) median and mel-filtered spectrogram, g) first difference, h) ROC

With 5-band ROCs of the median filtered squared magnitude spectrum, spectral transitions were indicated at least in one of the bands. However, the results also showed spurious peaks corresponding to noise and periodic spectral variations along the time axis, mainly in voiced speech segments, resulting in increased false detection. Spurious peaks were reduced by smoothing the magnitude spectrum along the time axis using a longer duration (typically 10 ms) median/mean filter, but at the cost of masking short duration spectral transitions and reducing the temporal accuracy of indicated transition points. We found out that the ROC of mel-filtered magnitude spectrum gives relatively better indication of spectral transitions and we studied it in detail for improved landmark detection.

5.3 Detection of Spectral Transitions Using Mel-filtered Magnitude Spectrum

Most of the landmark detection algorithms in the literature have been developed for applications in speech recognition. They depend on the analysis of speech signal in specific frequency bands which are related to spectral prominences of target sounds. Using multiple ROC functions (or other detection parameters) from specific bands can improve detection performance by tracing spectral transitions occurring in different bands of the spectrum. However, boundaries of the spectral transitions are not fixed. The locations and bandwidths for a single phoneme vary across speakers and speaking conditions. Instantaneous pitch variations within voiced phonemes can result in transition from one band to another of the spectral prominence of the phonemes, which would then be detected as landmarks. Therefore, choice of fixed frequency bands limits the performance. The ROC of median filtered magnitude spectrum discussed earlier indicates instants of salient signal transitions such as voicing onsets. However, it fails to clearly show gradual spectral changes such as voicing offsets and may lead to the detection of false landmarks due to harmonic transitions.

In order to address these problems, we consider filtering of the magnitude spectrum along the frequency axis using a bank of filters whose bandwidth varies according to the human perception scale. Spectral transitions convey perceptual information important for understanding the different phonemic and sub-phonemic sounds [43]. Therefore, critical band filtering of magnitude spectrum is likely to improve detection performance by emphasising perceptually significant spectral components and suppressing harmonic transitions. Mel-filter bank with number of filters equal to the number of FFT points has been selected for this purpose. ROC contour of the mel-filtered (along frequency) and median smoothed (along the time axis) magnitude spectrum was computed as explained in the previous section, for locating the spectral transitions. Implementation of the mel-filter bank and spectral smoothing along the time axis are discussed in the following sections.

5.3.1 Mel-filter bank

Mel-scale approximately maps the perceived frequency of a pitch onto the linear (acoustic) frequency scale. Mathematically, this mapping is given by [30]

$$\text{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{5.1}$$

where f is the linear-scale frequency in Hz and mel(f) is the mel-scale frequency. The relation is approximately linear for frequencies below 1 kHz and logarithmic for frequencies above 1 kHz. A Mel-filter bank is designed to mimic the auditory filter which is approximated as a bank of band pass filters with nearly constant quality factor [1], and hence it captures the perceptually important frequency components of speech spectrum [12], [24], [30]. The bandwidths of the individual filters are chosen to obtain high frequency resolution at lower frequencies to which our hearing system is highly sensitive and low resolution at high frequencies. Mel-filtering is often used for extracting mel-frequency cepstral coefficients for speech and speaker recognition systems [2], [16], [30], [47]. We have used a bank of closely spaced mel-filters for smoothing harmonic structures and clicks along the frequency axis, and to accentuate weak and/or gradual spectral transitions such as vowel ends.

In designing the filter bank, the frequency samples f_k of N-point DFT were taken to be the center frequencies in the linear scale and they were converted to the mel-scale using

$$\text{mel}(f_k) = 2595 \log\left(1 + \frac{f_k}{700}\right) \tag{5.2}$$

where $f_k = kF_s/N$, where k is the FFT index and F_s is the sampling frequency. The lower and upper cut-off frequencies for each filter in the mel-scale were obtained by selecting a uniform mel bandwidth, corresponding to the linear-scale bandwidth B_L of the lowest mel-filter. The bandwidth B_L was chosen to be about twice the pitch frequency in order to mask pitch harmonics while maintaining formant transitions. Different values of B_L between 50 and 500 Hz were used. While narrowband filters did not mask the harmonic structure, large bandwidth filters (B_L > 400 Hz) masked formant transitions and resulted in wider first difference bars around signal boundaries. The results were less sensitive to bandwidth values between 100 and 300 Hz, and B_L = 200 Hz was used for further investigation. For each filter with linear scale center frequency f_C, the mel-scale lower mel(f_L) and upper mel(f_U) cut-off frequencies were obtained by

$$\begin{aligned} \text{mel}(f_L) &= \text{mel}(f_C) - \text{mel}(B_L)/2 \\ \text{mel}(f_U) &= \text{mel}(f_C) + \text{mel}(B_L)/2 \end{aligned} \tag{5.3}$$

The linear scale lower cut-off f_L and upper cut-off f_U values were computed using the inverse relation, given as

$$f_k = 700\left(10^{\text{mel}(f_k)/2595} - 1\right) \tag{5.4}$$

The bandwidth B of each filter is linearly related to the center frequency, which can be derived as follows:

$$\begin{aligned} B &= f_U - f_L \\ &= 700\left(1+\frac{f_C}{700}\right)\left(\sqrt{1+\frac{B_L}{700}}-\left(\sqrt{1+\frac{B_L}{700}}\right)^{-1}\right) \end{aligned} \tag{5.5}$$

Figure 5.4 a) shows the linear-scale upper and lower cut-off frequencies and bandwidth as function of the linear frequency. The magnitude response of the mel-filters is triangular. For each filter with center frequency index k_C, lower cut-off index k_L and upper cut-off index k_U, the magnitude response was calculated using the formula

$$H_k(m) = \begin{cases} \dfrac{m-k_L}{k_C-k_L}, & k_L \le m < k_C \\ \dfrac{k_U-m}{k_U-k_C}, & k_C \le m < k_U \\ 0, m < k_L, & m > k_U \end{cases} \tag{5.6}$$

A part of the magnitude response of the triangular mel-filter bank is given in Fig 5.4 b).

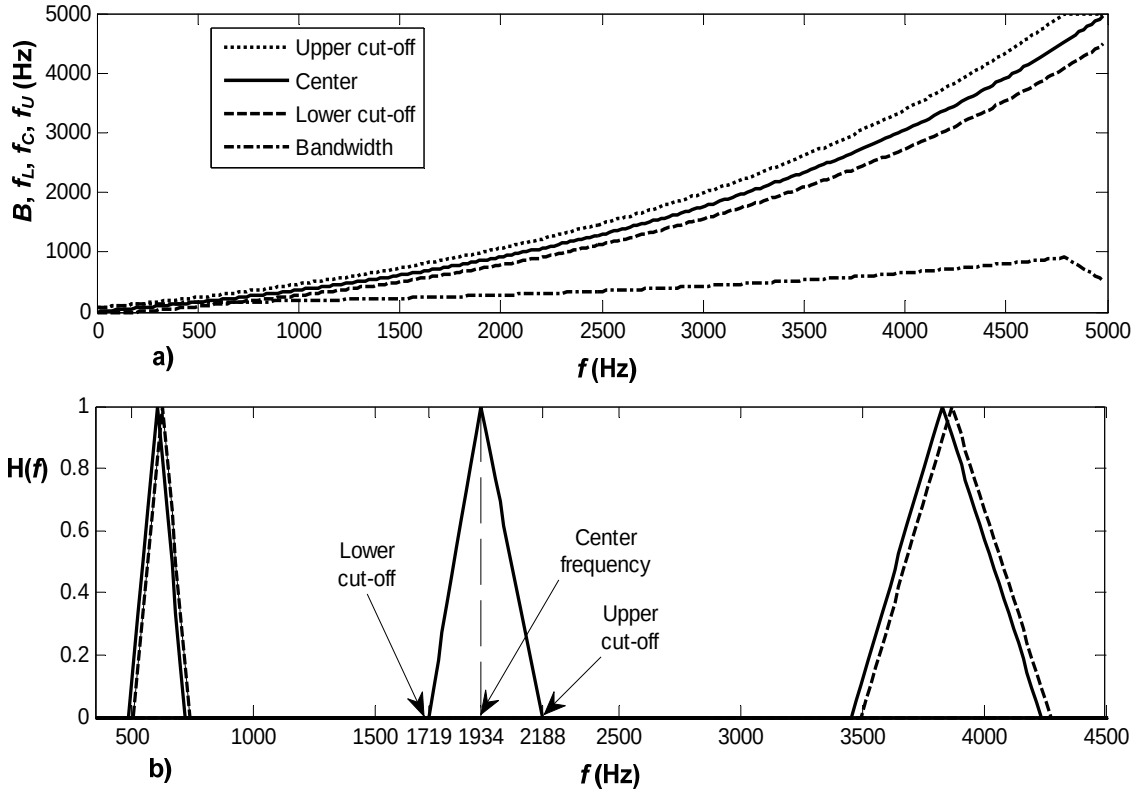

Fig. 5.4 Typical mel-filter bank characteristics: a) mel warped filter parameters in the linear scale: (- - -) bandwidth B (Hz), (—) upper cut-off frequency f_C (Hz) and (· · ·) lower cut-off frequency f_L (Hz, b) magnitude response of a part of the mel-filter bank

5.3.2 Spectral Smoothing in Time

Short-time magnitude spectrum of speech signal shows random intensity variations along the time axis. Particularly, the spectrogram of voiced segments obtained using a short-time window displays alternating dark and light vertical regions that appear as significant transitions. As a result, the first difference of the magnitude spectrum gives multiple high intensity values within the duration of the phonemes. Using a long duration window (one that includes several pitch periods) can smooth periodic spectral variations and noise; however, it results in poor time resolution. In order to address this problem, averaging and median smoothing of the mel-filtered magnitude spectrum along the time-axis were investigated separately.

Spectral averaging in time significantly minimizes spurious peaks due to periodic spectral variations in voiced segments. The 2-dimensional plot of the first difference of the averaged squared magnitude spectrum shows distinct boundaries at signal transitions. However, averaging is likely to cause the following problems: 1) it could result in backward and forward smearing of the signal resulting in timing misalignment of the detected spectral transitions and the actual transitions; 2) clicks and small signal variations could be widened in duration and detected as landmarks, and 3) low energy bursts may get masked by surrounding high intensity signal regions (e.g. vowels) and may not get detected. These problems were partly minimized by a careful choice of the averaging window length and the time shift between successive frames. Median smoothing of the squared magnitude spectrum was also examined in order to reduce the timing misalignment problem. A median filter has the useful property of suppressing low level impulse-type noise while preserving significant signal discontinuities. However, short-time median filter may not be effective in removing spectral variations between pitch periods.

In both cases, the length of the filter has a significant effect on the output. A longer duration (narrowband) filter can be effective in removing spectral variations, but it reduces the temporal accuracy in locating spectral changes. It results in masking of short spectral transitions and shifting of detected landmarks. A short duration (wideband) filter, on the other hand, preserves the fine temporal resolution but it is less effective in smoothing individual pitch period transitions and noise.

5.3.3 Detection of Spectral Transitions

Figure 5.5 shows the block diagram of the landmark detector using mel-filtered squared magnitude spectrum. The input speech signal was sampled at 10 kHz and 512-point FFT was computed using 6 ms Hanning window every 3 ms. The square of the magnitude spectrum was filtered by a 256-triangular mel-filter bank along the frequency axis. The mel-filtered squared magnitude spectrum $\hat{S}_n^2(k)$ for n^{th} frame is given by

$$\hat{S}_n^2(k) = \sum_{m=1}^{N/2} \left|S_n(m)\right|^2 \left|H_k(m)\right|, \quad k = 1,...,256 \tag{5.7}$$

where N is the FFT length, k and m are frequency indices, $S_n(k)$ represents FFT of the input samples and $H_k(m)$ is the magnitude response of the k^{th} filter. Varying the number of mel-filters between 256 and 32 gave similar ROC contours, but performance of the voicing detector discussed later in this section, improved with increasing the number of filters. The output of the mel-filter was given to two separate detectors based on the type of spectral smoothing used along the time axis. One uses average of the mel-filtered squared magnitude spectrum computed from $2M$+1 neighbouring samples as

$$\tilde{S}_n^2(k) = \frac{1}{2M+1} \sum_{m=n-M}^{n+M} \left|\hat{S}_n^2(k)\right|^2 \tag{5.8}$$

where $\tilde{S}_n^2(k)$ is the output of the filter. The other uses median filtering in which the output is given as

$$\tilde{S}_n^2(k) = \text{median}\left\{\hat{S}_n^2(k-M), \hat{S}_n^2(k-M-1), ..., \hat{S}_n^2(k+M)\right\} \tag{5.9}$$

The choice of the averaging and median window length (M) was based on observing the effects of different values, and M = 2 for averaging and M = 5 for median filtering were typically used.

The first difference $D_n(k)$ for n^{th} frame was computed from the log of the filtered squared magnitude spectra as

$$D_n(k) = \log|\tilde{S}_n^2(k)| - \log|\tilde{S}_{n-n_0}^2(k)| \tag{5.10}$$

where n_0 is the difference step in frames. The logarithm operation is applied to avoid the need for normalization for overall amplitude variation and compress small spectral variations. Different values of the window length, window shift and difference step (n_0), were tested to study their effects and to select suitable values. Voiced sounds required larger values of window length, window shift and difference step (n_0) to smooth out the spectral variations between quasiperiodic cycles and to locate voicing onset and offset boundaries using first difference. However, this had the effect of masking short duration bursts and reducing the temporal accuracy. A window shift of 3 ms and difference step of 6 ms gave a good trade-off between spectral smoothing and temporal accuracy and were used for our results.

The sum of the first 50% largest absolute first difference values in each frame was taken as a rate of change (ROC) parameter for landmark search. This ROC has distinct peaks at prominent signal transitions, but it can also give several closely spaced peaks during non-abrupt transitions. In order to suppress the smaller peaks and to simplify the peak picking algorithm, the ROC was smoothed using stationary wavelet transform (SWT) denoising [48], [49]. SWT is a time-invariant wavelet transform in which the transform coefficients at a given decomposition

level are obtained by convolving the output of the immediate lower level with the upsampled version of the filter coefficients. SWT denoising is used to suppress spurious peaks and to emphasize the prominent peaks without disturbing their temporal locations. In this step, a 3-level SWT decomposition of the signal was computed using Daubechies1 (Haar) wavelets and the detail coefficients in each level were smoothed by soft thresholding. The signal was then reconstructed from the approximation and modified detail coefficients using the inverse transform.

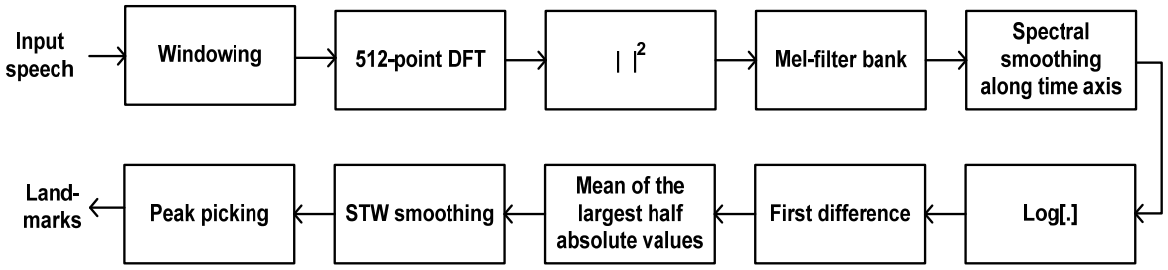

Fig. 5.5 Detection of spectral transitions using mel-filtered squared magnitude spectrum

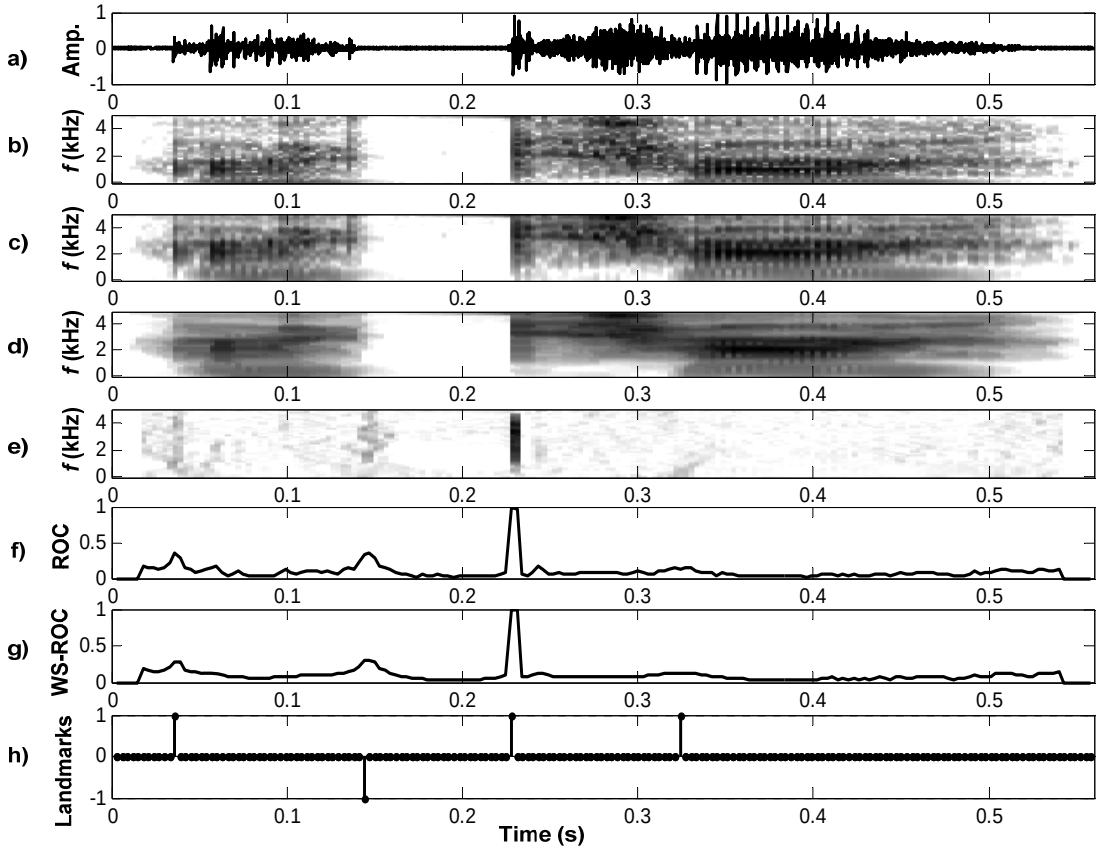

Fig.5. 6 Landmark detection for utterance /aka/ from female speaker FS6: a) waveform, b) spectrogram, c) mel-filtered spectrogram, d) median and mel-filtered spectrogram, e) first difference of median and mel-filtered spectrogram, f) ROC, g) wavelet smoothed ROC h) landmarks

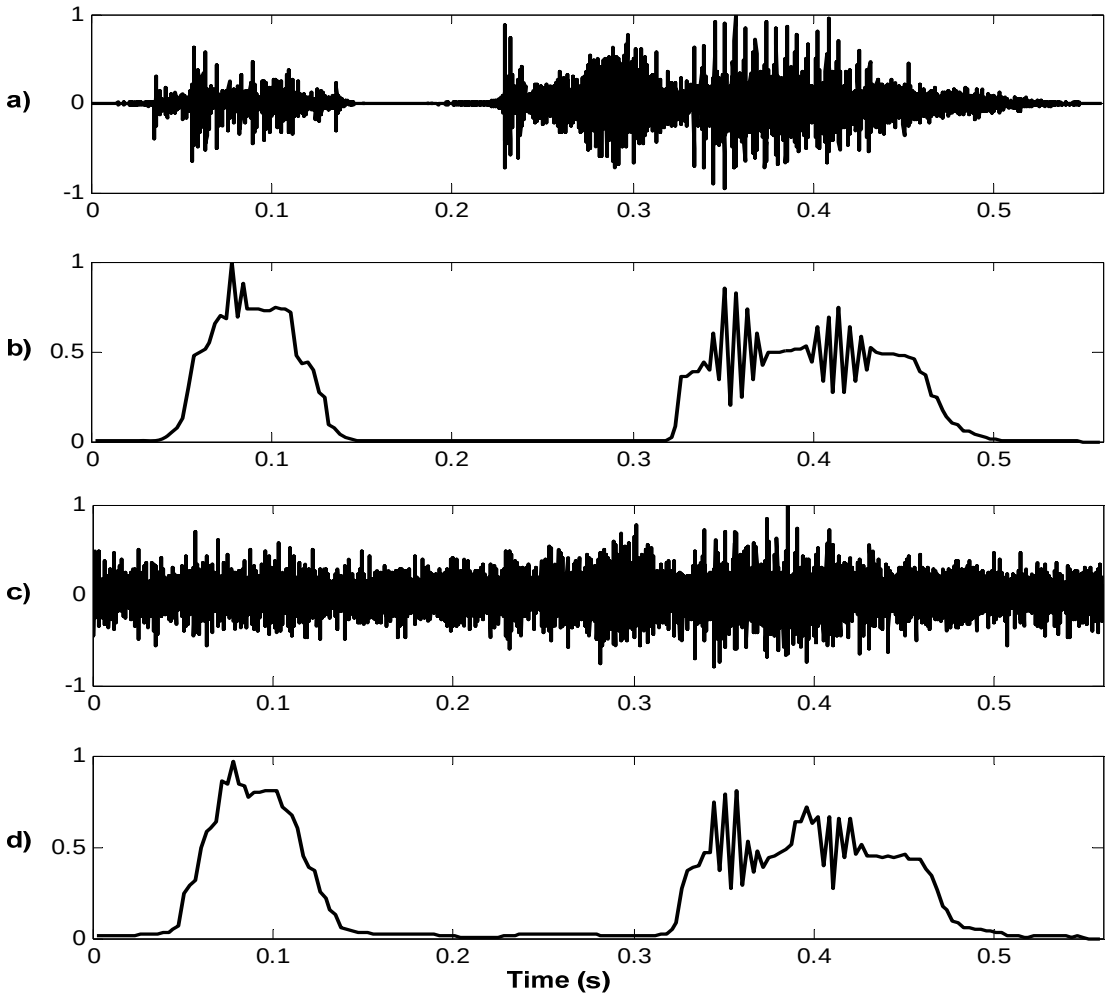

Fig. 5.7 Voicing detection results for utterance /aka/ from female speaker FS6: a) speech waveform, b) spectral slope c) speech waveform with AWGN of 5 dB SNR, d) spectral slope of noise corrupted speech

During peak picking, the smoothed ROC signal was segmented into frames of 300 ms (maximum vowel duration) and all peaks above a local threshold in each frame were taken as potential landmarks. First difference of 30 ms averaged energy computed with a time step of 27 ms centered at the time of interest was used to determine whether a peak represented falling or rising transition. A peak was decided as a falling transition (closure) if the first difference at the peak location is negative and rising (burst) otherwise. Phonetic duration constraints such as minimum stop closure, frication and voicing onset/offset durations were imposed on the selected peaks to detect the final spectral transitions. Based on observation of temporal characteristics of different VCV utterances, the minimum duration values were chosen to be 6 ms between successive positive peaks (which corresponds to minimum voicing onset and/or frication durations), 10 ms between consecutive negative peaks (minimum voicing offset) and 15 ms between successive positive and negative peaks (minimum closure/silence duration). An

illustration of the detection process for typical VCV utterances, with results at each stage, is shown in Fig. 5.6.

Voicing onset/offset points were detected based on first difference of the mel- and median filtered squared magnitude spectrum along frequency, we describe as the spectral slope. For voiced sounds, the magnitude spectrum has strong values around harmonics and formants, and weak values elsewhere; hence taking the first difference along the frequency axis results in high intensity values in the lower frequency region. On the other hand, unvoiced sounds have almost uniform spectrum and hence the first difference gives small values in the entire spectrum. The spectral first difference at each sampling time n, $a_n(k)$, was calculated as

$$a_n(k) = |\, \widetilde{S}_n^2(k + k_0) - \widetilde{S}_n^2(k)\,|, 1 \le k \le N/2 \tag{5.11}$$

where k_0 is difference frequency step. Several values of difference frequency step were tried out and the results were almost the same for difference steps below 400 Hz. A frequency step of 2 samples (40 Hz) was used for our analysis. The spectral slope parameter was computed as the sum of the largest half of the difference values in each frame, and its normalized value was used for voicing onset/offset detection based on a selected threshold. A value was set to 1 if it was above the threshold and to 0 otherwise. First difference of the resulting signal was computed between successive samples in time, and positive and negative peaks indicated voicing onset and offset, respectively. The voicing detector was tested on VCV utterances corrupted with additive white Gaussian noise of different signal-to-noise ratio (SNR). The spectral slope was found to have robust performance for SNR down to 5 dB. Figure 7 shows the spectral slope for utterance |aka| from female speaker FS6, corrupted by noise at 5 dB SNR.

5.4 Detection of Stop Landmarks Using Mel-filtered Spectrum

Stop consonants typically consist of a silence interval (except for word initial stops) due to complete closure of vocal tract, followed by a short frication burst corresponding to the abrupt release of vocal tract occlusion. Durations of the closure and burst segments are context and speaker-dependent, and vary from 5 ms to about 100 ms. Measurement of the closure and burst durations for speaker recognition requires detection of the closure, and burst onset and offset instants with high temporal accuracy. Here, rate-of-change (ROC) of mel-filtered magnitude spectrum is used to detect the abrupt transitions in speech signal, and additional spectral and durational properties are used to indentify stop landmarks. Two types of stop landmark detectors were studied; for reference, we will call them landmark detector-1 (LMD1) and landmark detector-2 (LMD2). They differ in the method of picking the ROC peaks and validating the selected peaks as stop landmarks. In LMD1, all ROC peaks greater than a local threshold are picked as potential landmarks. Voicing detection, spectral flatness measure and intensity level

parameters are used as additional features to select stop landmarks. In LMD2, closure segments are located based on a product of spectral flatness measure and log energy, and ROC peak picking is carried out only around the start and end of the closures. Spectral flatness, log energy and their product are used for selecting the final landmarks. The following section discusses the detection process using LMD1 and evaluation results are presented in sections 5.6. Explanation of LMD2 and detection results are given in sections 5.7 and 5.8.

5.5 Landmark Detector LMD1

Spectral transitions are detected using the technique discussed in Section 5.3.3. Magnitude spectrum of the input speech, sampled at 10 kHz, was computed using 6 ms Hanning window and 512-point DFT every 3 ms. The squared magnitude spectrum was filtered along the frequency axis by 256-triangular mel-filter bank with lowest filter bandwidth of 136 Hz. Each sample is obtained as a perceptually weighted sum of its neighbours. The purpose of this filtering is to suppress the harmonic structure without loosing the formant structure. The output of each mel-filter was further median filtered along the time axis. For sentences, an 11-point filter was used to smooth out amplitude variations between pitch periods and other undesired spectral transitions without significantly affecting the actual transitions. A 7-point filter was found to be sufficient for the same task for VCV utterances which have relatively longer segmental durations, and more distinct stop closure and burst transitions. First difference of the log spectra was computed using (5.10) with time step of 6 ms. A fast frame rate (3 ms) and small difference step are required to track the abrupt closure and burst transitions. A ROC function was calculated as the mean of the 50% largest absolute difference values in each frame, smoothed by 3-level SWT using Haar wavelet. A silence-burst transition in stops results in large difference values along the entire spectrum, compared to other transitions that occur in specific bands; hence the ROC parameter gives relatively distinct peaks for stop bursts.

Additional parameters such as voicing, spectral flatness measure and its slope, and sum of magnitude spectra, were used to validate the detected transitions as stop landmarks. The spectral flatness and intensity measure parameters were obtained from the FFT $S_n(k)$ of the input samples, computed using 20 ms Hanning window and 512-point DFT every 1 ms. The relatively longer window is needed for accurately estimating the spectrum flatness, while the high frame rate is needed to capture the short duration transitions. The spectrum flatness was measured in terms of Wiener entropy, $E_W(n)$, of the magnitude spectrum, calculated as [44]

$$E_W(n) = \sum_{k=1}^{N/2} \log|S_n(k)| - \log\left(\sum_{k=1}^{N/2} |S_n(k)|\right) \tag{5.11}$$

where n is the time (frame) index, N is the number of DFT points and $S_n(k)$ is the k^{th} FFT sample at frame n. $E_W(n)$ is used to separate voiced and unvoiced regions. It has large values (near unity) for unvoiced segments and, low and peaky values for voiced sounds. Fricative and nasal bursts, and burst like sounds after pauses have flat spectrum similar to stops and qualify the spectral flatness measure test. Two additional parameters were used to identify stop closures and bursts: 1) slope of the spectral flatness measure and, 2) measure of sum of magnitude spectra as measure of intensity level, calculated as

$$S_I(n) = \sum_{k=1}^{N/2} |S_n(k)| \tag{5.12}$$

where N = 512. Production of stop consonants often involves complete closure of the vocal folds before the burst and their closure interval has very low intensity as compared to that of most fricatives and nasals. The closure-burst transitions in stops are generally found to be more abrupt than that of fricatives, nasals and pauses. Hence Wiener entropy for stops has peaks with large slopes around bursts, which were captured from the first difference between successive frames.

A voicing detector based on spectral slope of the mel and median filtered magnitude spectrum, computed using 6 ms Hanning window and 512-point DFT every 3 ms, was used to detect frication offset landmarks missed by the ROC parameter and to remove ROC peaks within voiced regions. First difference of the mel and median filtered magnitude spectrum was computed with 2 frequency samples (40 Hz) step. A spectral slope computed as the normalized mean of the 25% largest absolute difference values in each frame, was used as parameter for detecting voiced segments. This parameter has large values for voiced sounds which have a well-defined harmonic and formant structure, and small values for unvoiced sounds which have almost flat spectrum. A continuous segment of minimum duration 40 ms and having spectral slope greater than 0.0125 (empirically determined value) was decided to be voiced, and its initial and final points were labelled as voicing onset and voicing offset landmarks respectively.

Spectral transitions indicating start of closure (-p) and release (+p) were detected from the ROC parameter using a peak picking method with local threshold and average energy, as discussed in Section 5.3.3. All +p peaks with $E_W(n)$ greater than 0.5 and amplitude normalized absolute value of slope of $E_W(n)$ above 0.45 within the preceding 30 frames were taken to be valid stop burst and frication offset landmarks. Similarly, -p peaks with $E_W(n)$ greater than 0.5, normalized absolute value of slope of $E_W(n)$ above 0.45 and $S_I(n)$ below 0.025 within the succeeding 30 frames were taken as valid stop closure landmarks. In addition, +p and -p peaks neighboured by $E_W(n)$ greater than 0.65 within the specified time margin were considered as stop landmarks. A voicing onset mark preceded by $E_W(n)$ greater than 0.5 and normalized absolute slope of $E_W(n)$ above 0.45 was detected as a stop frication offset. Remaining voicing onset and

offset points, and any +p and -p peaks inside a voiced segment were eliminated. Finally, minimum and maximum duration (10 ms and 80 ms for closure, and 10 ms and 50 ms for frication) constraints were imposed. The slope of Wiener entropy was found effective in removing fricative bursts, and most of the insertions were due to nasals and short pauses.

5.6 Results Using LMD1

The landmark detection method was evaluated on VCV utterances and TIMIT sentences. The results are presented in the following two sections.

5.6.1 Detection on VCV Utterances Using LMD1

The method was tested on 18 VCV syllables formed using 6 stop consonants (/b/, /d/, /g/, /p/, /t/, /k/) and 3 vowels (/a/, /i/, /u/), spoken by 5 female and 5 male speakers. Closure, burst and voicing onset landmarks were detected from ROC (along time) of mel-filtered magnitude spectrum using voicing detection as a guide. In this case, the ROC contour was computed as the product of the arithmetic mean and the geometric mean of the largest 50% absolute values of the first difference (along time axis) of the mel and median filtered squared magnitude spectrum. Voicing onset and offset points were located from the spectral slope of mel and median filtered magnitude. The highest peak in the ROC parameter located between the voicing offset of the first vowel and onset of the second vowel was taken to be the burst landmark. A ROC peak in the neighbourhood of a detected voicing offset and at least 15 ms before the burst location was assumed to be the closure landmark. The actual voicing onset landmark was taken to be the ROC peak around a detected voicing onset and at least 6 ms after the burst landmark. In case voicing offset or onset is missed by the voicing detector, the highest ROC peak located after 90 ms (minimum vowel + closure duration) from the start of the syllable and at least 50 ms before the end, was taken to be the burst landmark. A peak located 15 ms to 60 ms before the burst was considered as closure landmark and a peak 6 ms to 40 ms after the burst was detected as the voicing onset. A landmark was assumed missed if it was located outside the 30 ms neighbourhood of the respective manual landmark, while a detected landmark which was found more than 30 ms away from any manual landmark was counted as insertion.

Figure 5.8 shows stop landmark detection for /iti/ spoken by female speaker FS6, using ROC of squared magnitude spectrum filtered by a 256-triangular mel-filter bank along the frequency axis and 7-point median filter along the time axis. The numbers of detected landmarks (out of 180 landmarks of each type) at 6 levels of temporal accuracy relative to manually labelled landmarks are given in Table 5.1. Figure 5.9 shows the detection rates. Stop burst landmarks were detected with rates of 52%, 85%, 93% and 95% within 3, 5, 10, and 15 ms interval of the manually located landmarks, respectively. The detection rates for closure landmarks were 62%,

74%, 87%, 92%, 96%, and 97% within 3, 5, 10, 15, 20 and 30 ms respectively. Voicing onsets were detected at 45%, 66%, 91%, 97%, 98% and 99% with temporal accuracies of 3, 5, 10, 15, 20 and 30 ms respectively. The high detection accuracy (above 90% at 10 ms accuracy) shows that ROC of mel-filtered magnitude spectrum obtained at high frame rate tracks the abrupt spectral transitions in the speech signal. There were insertions mainly due to strong clicks in the silence and frication regions of the syllables which resulted in a few missed burst landmarks. The overall detection rates were 53%, 75%, 90%, 95%,

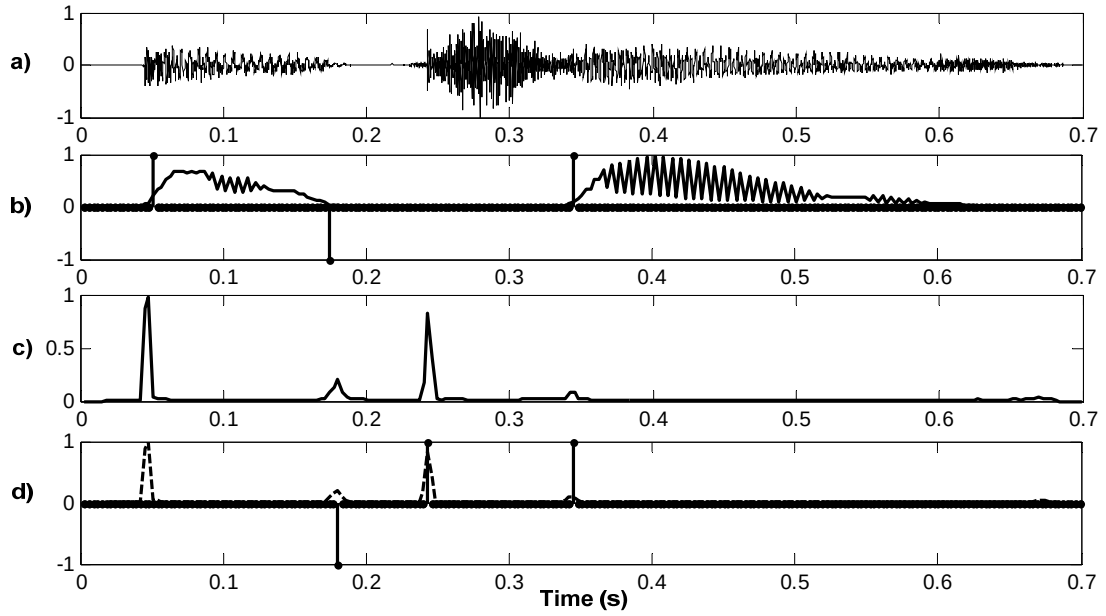

Fig. 5.8 Stop landmark detection for utterance /iti/ from female speaker FS6 using LMD1: a) waveform, b) spectral slope, c) ROC, e) detected stop landamrks

Table 5.1 Number of detected stop landmarks in 180 VCVs from 10 speakers (5 male, 5 female) using LMD1

Speaker	landmark	Detection accuracy relative to manual						Insertions
		<= 3	<= 5	<= 10	<= 15	<= 20	<= 30	
Female	Closure	57	68	79	83	89	90	0
	Burst	58	79	86	87	87	87	3
	VOT	34	62	82	87	88	89	1
	Total	149	209	247	257	264	266	4
Male	Closure	56	66	79	84	84	85	5
	Burst	35	74	82	84	84	84	6
	VOT	47	58	82	89	90	90	0
	Total	138	198	243	257	258	259	11
Overall	Closure	113	134	158	167	173	175	5
	Burst	93	153	168	171	171	171	9
	VOT	81	120	164	176	178	179	1
	Total	287	407	490	514	522	525	15

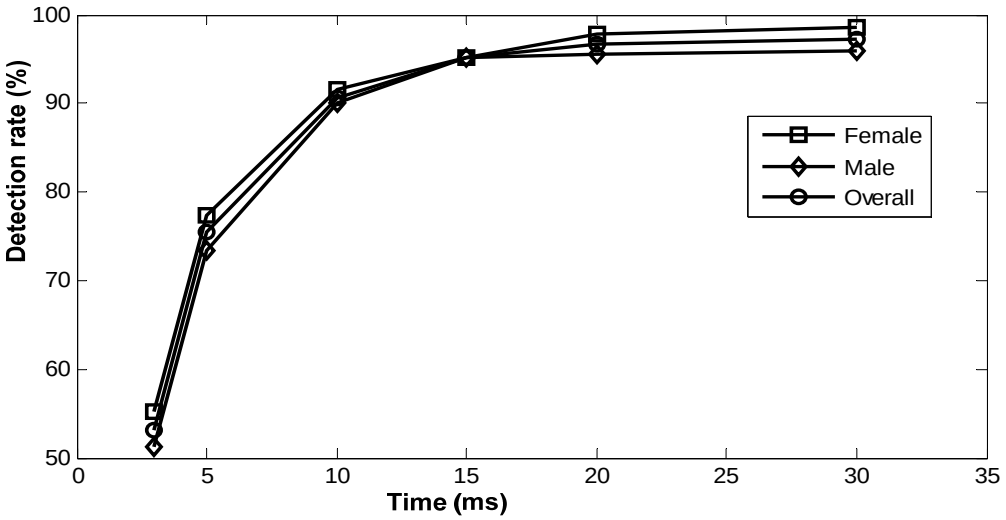

Fig. 5.9 Detection rates of stop landmarks in VCV syllables for different temporal accuracies, using LMD1

96%, and 97% within 3, 5, 10, 15, 20 and 30 ms respectively of the manually labelled landmarks, and 3% insertion. Compared to Jayan's stop landmark detection results on the same set of VCVs using Gaussian mixture modelling of speech spectrum [46], our method has more than 35%, 20%, and 20 improvement in the detection rate of closure landmarks with 5, 10 and 15 ms accuracy respectively. However, the detection rates at 5 ms for burst and voicing onset landmarks are lower in our method by 5% and 10% respectively.

The results were analysed to see if the performance of the method varies gender wise. There is slight difference in terms of temporal accuracy and overall detection rates between the two set of speakers. The results are higher for female speakers. The detection rates for the female speakers were 55%, 77% and 98.5% within 3, 5 and 30 ms of the manual landmarks, respectively. The corresponding values for the male speakers were 51%, 73% and 96%. Such gender and speaker-dependent variations are expected, particularly for landmark detection methods based on direct spectral features, because different speakers have different spectral characteristics (such as pitch and formants) and speaking rates, which require a different set of analysis windows and frame rates to achieve desired temporal and spectral resolution. The window length and frame rate values can be tuned for detecting specific class of sound, but the performance will still vary with speakers and contexts.

5.6.2 Detection on TIMIT Sentences Using LMD1

A test database consisting of 50 TIMIT sentences from 3 female and 2 male speakers was used for evaluation. Out of the 10 sentences spoken by each speaker, two were common for all the speakers. Manual transcriptions of stops (labelled in the TIMIT database as |b|, |d|, |g|, |p|, |t|, |k|,

|bcl,| |dcl|, |gcl|, |pcl|, |kcl|, |tcl|) were obatnied for comparison. Figure 5.10 shows stop landmark detection for the sentence "His captain was thin and haggard and his beautiful boots were worn and shabby." from female speaker FAKS0. Speaker-wise and overall detection rates of stop closure, burst and frication offset landmarks at 6 levels of temporal accuracy are given in Table 5.2. Table 5.3 gives the number of insertions. The overall detection rates are shown in Fig. 5.11.

Burst landmarks were detected at improved temporal accuracy, with rates of 72%, 84%, 89%, 93%, 94% and 95% respectively within 3, 5, 10, 15, 20 and 30 ms of the manual landmarks. Detection rates for closure landmark were 36%, 57%, 77%, 85%, 89% and 93% at temporal accuracies of 3, 5, 10, 15, 20 and 30 ms respectively. Frication offset landmarks were detected with rates of 50%, 65%, 74%, 78%, 81%, 86%. There were insertions (55 closures, 59 bursts and 25 frication offsets) due to clicks, glottal stop, affricates, and transitions from pauses, nasals and fricatives to vowels and semivowels, and from pauses to nasals. Fricative insertions were reduced by the spectral flatness slope and intensity measures, and most of the insertions were mainly due to nasals and short pauses. The results show significant differences in detection rates among the 5 speakers. Since 8 out of the 10 sentences were different for each speaker, the results suggest that there could also be high context dependency. Overall, stop landmarks were detected at rates of 52%, 68%, 79%, 84%, 87% and 91% respectively within 3, 5, 10, 15, 20 and 30 ms of the manual landmarks, and insertion rate of 18.7%.

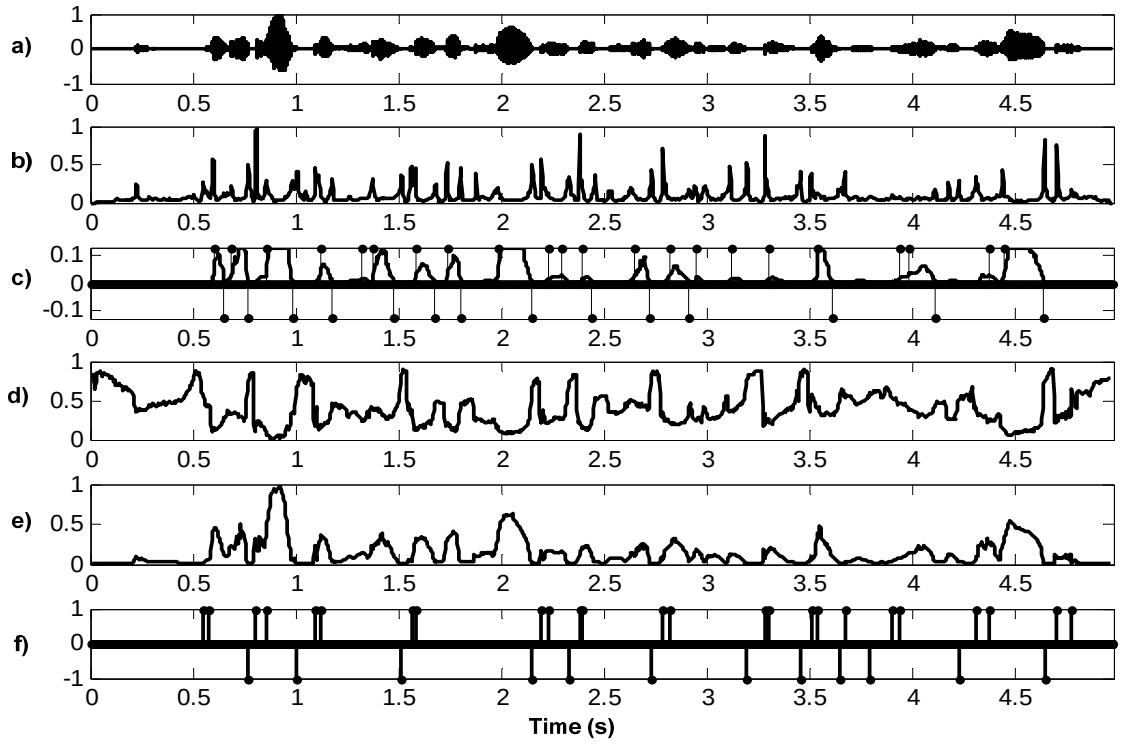

Fig. 5.10 Detection of stop landmarks using LMD1 for the TIMIT sentence "His captain was thin and haggard and his beautiful boots were worn and shabby." by female speaker FAKS0: a) waveform, b) ROC, c) spectral slope, d) $E_W(n)$, e) $S_I(n)$, f) detected stop landmarks

Table 5.2 Detection rates (%) of stop landmarks (b, d, g, p, t, k) in TIMIT sentences using LMD1

Landmark type	Speaker	Detection accuracy relative to manual transcriptions (in ms)						Insert-ions
		<=3	<=5	<=10	<=15	<=20	<=30	
Closures	FAKS0	36.8	66.7	82.5	89.5	94.7	96.5	8.8
	FDAC1	22.9	45.8	64.6	79.2	87.5	93.8	31.3
	FELC0	33.3	55.6	75.9	81.5	81.5	90.7	13.0
	MSKT0	44.7	61.7	78.7	83.0	89.4	89.4	31.9
	MWBT0	44.0	56.0	82.0	90.0	94.0	96.0	26.0
	Overall	36.3	57.4	77.0	84.8	89.5	93.4	21.5
Bursts	FAKS0	79.0	85.5	91.9	91.9	91.9	93.5	11.3
	FDAC1	67.9	81.1	83.0	88.7	90.6	94.3	30.2
	FELC0	67.3	76.4	81.8	83.6	89.1	90.9	14.5
	MSKT0	66.0	82.0	86.0	86.0	86.0	92.0	30.0
	MWBT0	56.0	74.0	84.0	84.0	84.0	98.0	26.0
	Overall	67.8	80.0	85.6	87.0	88.5	93.7	21.9
Frication offsets	FAKS0	54.7	67.9	79.2	81.1	83.0	84.9	5.7
	FDAC1	39.5	52.6	63.2	76.3	81.6	84.2	15.8
	FELC0	50.0	65.9	70.5	70.5	75.0	81.8	6.8
	MSKT0	60.5	73.7	78.9	78.9	81.6	92.1	10.5
	MWBT0	45.7	65.2	76.1	80.4	82.6	89.1	19.6
	Overall	50.2	65.3	74.0	77.6	80.8	86.3	11.4
All types	FAKS0	57.6	73.8	84.9	87.8	90.1	91.9	8.7
	FDAC1	44.6	61.2	71.2	82.0	87.1	91.4	26.6
	FELC0	50.3	66.0	76.5	79.1	82.4	88.2	11.8
	MSKT0	57.0	72.6	81.5	83.0	85.9	91.1	25.2
	MWBT0	48.6	65.1	80.8	87.7	89.7	94.5	24.0
	Overall	51.8	67.9	79.2	84.0	87.1	91.4	18.7

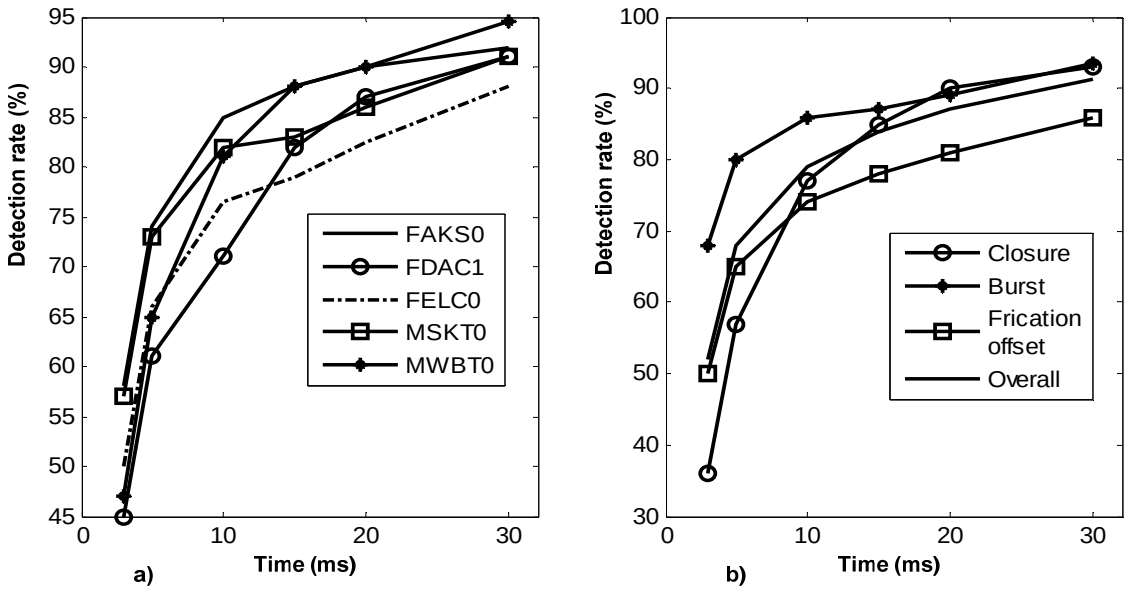

Fig. 5.11 Detection rates for TIMIT sentences using LMD1: a) speaker-wise rate, b) overall rates

Table 5.3 Number (percentage) of insertions in TIMIT sentences using LMD1

Insertion type	Number (percentage) of Insertions				Notes
	Closures	Bursts	Frication offsets	Total	
Pauses	12 (4.7)	12 (4.4)	6 (2.7)	30 (4)	Pause ('epi', and 'pau') preceding nasals, glides and fricatives,
Glottal stop	10 (3.9)	10 (3.7)	6 (2.7)	26 (3.5)	Brief closures denoted as 'q' in the TIMIT transcriptions
Nasals	8 (3.1)	8 (3.0)	3 (2.4)	19 (2.6)	Nasal('n', 'm', 'ng', 'nx') closures and bursts, esp. those preceded or followed by pauses and stop closures
Fricatives	12 (4.7)	13 (4.8)	5 (2.3)	30 (4.0)	3 ('f', 's', 'v') and 9 ('th','dh')
Semi-vowels	3 (1.2)	3 (1.1)	0 (0)	6 (0.8)	Closures in 'l', 'r', 'iy' followed by 'w'
Clicks	3 (1.2)	6 (2.2)	3 (1.4)	12 (1.6)	Clicks mostly in long pauses
Flaps	7 (2.7)	7 (2.6)	2 (0.9)	16 (2.1)	Sounds denoted by 'dx'
Total	55 (21.5)	59 (21.9)	25 (11.4)	139 (18.7)	55 closures and 25 frication. Flaps sound like stops, hence could be ignored, which reduces the total insertion rate to 16.6%

5.7 Landmark Detector LMD2

In the technique LMD1, discussed in Section 5.5, all ROC peaks above a local threshold were picked as potential landmarks, and only -p and +p peaks outside of voiced region that qualify the spectral flatness measure and intensity level tests were selected as stop landmarks. However, detection performance can be improved, both in terms of insertion rate and temporal accuracy, by looking for ROC peaks only in regions that manifest strong stop consonant features. We used closure and burst as the key features for picking stop landmarks from the ROC contour. Closures and bursts were detected based on Wiener entropy, log energy and their product. The Wiener entropy has large values in closure and frication regions, while log energy has large negative values during closures. We found out that the product of Wiener entropy and log energy gives a well-shaped waveform having large negative values only during closures, from which closure onset and offset landmarks can be detected with good temporal accuracy, using a simple threshold. Wiener entropy and its slope, maximum deep in the closure region indicated by the product function, maximum log energy around burst, and duration constraints were used to separate stop closures and bursts from fricatives, nasals and pauses.

The ROC function for detecting the spectral transitions was obtained using the method discussed in Section 5.3.3. Fourier transform of the input signal was computed using 512-point

DFT, 6 ms Hanning window and 3 ms frame rate. The squared magnitude spectrum was filtered using 256-triangular mel-filter bank along the frequency axis and 11-point median filter along the time axis. First difference of the resulting spectra was calculated using 6 ms time step. The ROC parameter was taken to be the mean of largest half of the absolute difference spectra, smoothed by 3-level Haar wavelet SWT. The Wiener entropy and log energy were obtained from the FFT $S_n(k)$ of the input speech, computed using 20 ms Hanning window and 512-poind DFT every 1 ms. The Wiener entropy $E_W(n)$ of the squared magnitude spectrum was calculated using (5.11). The log energy $E_L(n)$ was computed as

$$E_L(n) = \sum_{k=1}^{N/2} \log |S_n(k)|^2 \tag{5.12}$$

where N is the number of DFT points, and k and n are the frequency and time indices, respectively. The parameter used for detecting closure, $X_S(n)$, was given by

$$X_S(n) = E_W(n) E_L(n) \tag{5.13}$$

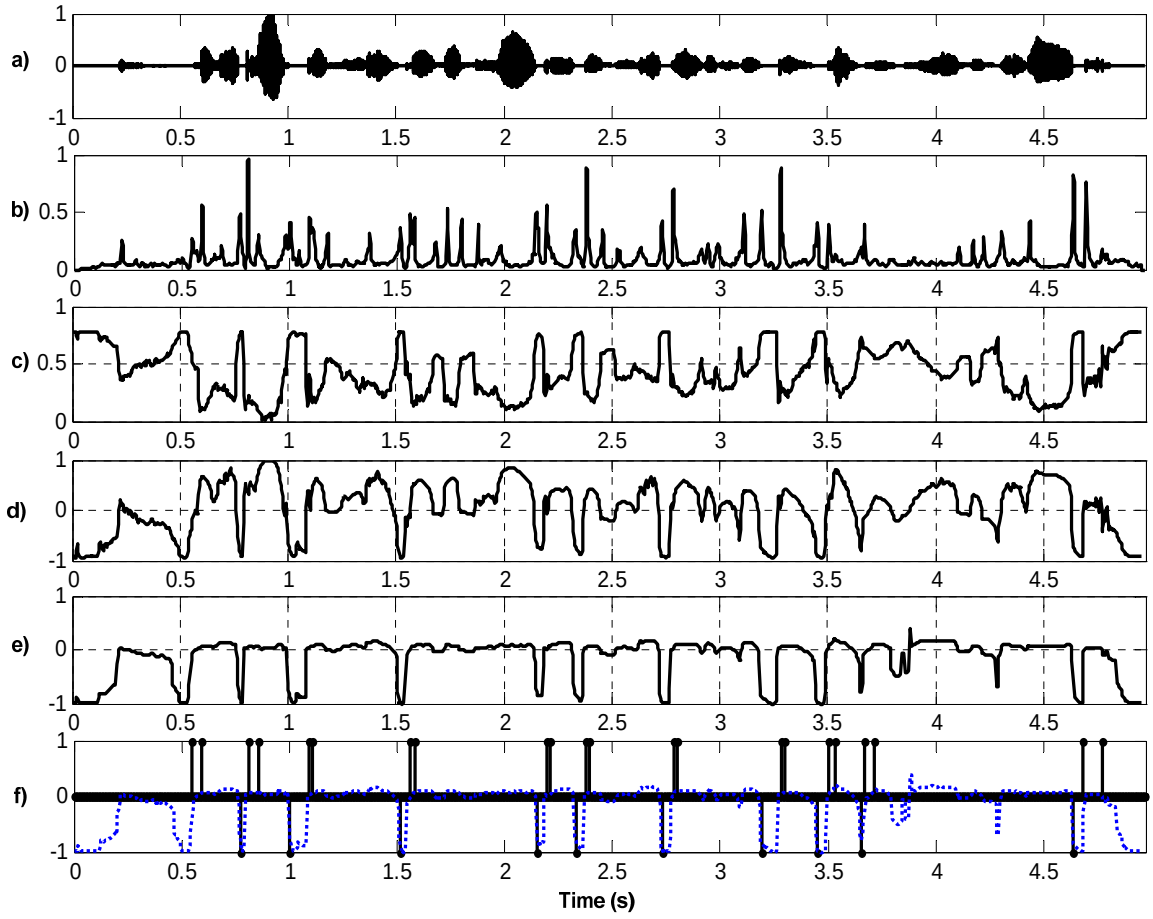

Fig. 5.12 Detection of stop landmarks using LMD2 for the TIMIT sentence "His captain was thin and haggard and his beautiful boots were worn and shabby." By female speaker FAKS0: a) waveform, b) ROC, c) $E_W(n)$, d) $E_L(n)$, e) $X_S(n)$, f) detected stop landmarks

Closure segments were detected from the amplitude normalized values of $X_S(n)$, $E_W(n)$ and its slope, and $E_L(n)$, using empirically determined thresholds and minimum duration constraints. A speech segment with $X_S(n)$ less than -0.225 for at least 12 ms was taken as a valid silence. Minimum separation between successive silences was taken to be 30 ms. A closure segment having within its interval minimum of $X_S(n)$ less than -0.5, maximum of $E_W(n)$ greater than 0.5, maximum of slope of $E_W(n)$ greater than 0.4, and maximum of $E_L(n)$ within 20 ms after the burst greater than 0.2, was detected as stop closure. The actual stop closure, burst and frication offset landmarks were detected by picking ROC peaks around the detected stop closure segments. The largest ROC peak within 10 ms neighbourhood of the start of a detected closure was taken to be a stop closure landmark, while the largest peak found within 10 ms interval of the end point of a closure and at least 10 ms after a detected closure landmark, was taken to be a stop burst landmark. A ROR peak located between 6 ms to 60 ms after a detected stop burst landmarks was taken to be the frication offset landmark.

5.8 Results Using LMD2

The method was evaluated on the same 50 TIMIT sentences used for evaluating LMD1. Figure 5.12 shows the ROC, Wiener entropy, log energy, product of Wiener entropy and log energy, and detected stop landmarks for the sentence "His captain was thin and haggard and his beautiful boots were worn and shabby," spoken by female speaker FAKS0. The speaker-wise and overall detection rates for each type of landmark at 6 level of temporal accuracy are given in Table 5.4. Table 5.5 shows the type and number of insertions. The overall detection rates for 5 speakers and the overall detection rates for the 3 types of stop landmarks are shown in Fig. 13 (a) and Fig. 13 (b) respectively. There were a total of 272 target stop consonants, with 256 closures, 272 bursts and 220 frication offsets landmarks to detect. The detection rates for closure landmarks were 40%, 59%, 80%, 87%, 90% and 93%, respectively within 3, 5, 10, 15, 20 and 30 ms of the manual landmarks. Stop burst landmarks were detected at high accuracy with rates of 59%, 78%, 88%, 91%, 94% and 95% at 3, 5, 10, 15, 20 and 30 ms temporal accuracy, while frication offset landmarks were detected at 53%, 68%, 79%, 83%, 86% and 90%. The high detection accuracies of closure and burst landmarks show the accuracy of the closure detector and the ability of the ROC function to distinctly show abrupt spectral transitions. The overall detection rates of all types of stop landmarks were 52%, 69%, 83%, 87%, 90% and 93%. There were a total of 48 insertions due to short pauses, fricatives, nasals, glottal stop and clicks. As before, there is significant difference in performance of the method for different speakers and contexts, especially in the detection rate with 3 ms accuracy and number of insertions. This could account for differences in the abruptness of closure and burst transitions, and duration of closures and frication segments among the speakers, which require the use of different framing and smoothing

Table 5.4 Detection rates of stop consonant landmarks in TIMIT sentences using LMD2

Landmark type	Speaker	Detection accuracy relative to manual transcriptions (in ms)						Insertion
		<=3	<=5	<=10	<=15	<=20	<=30	
Closures	FAKS0	38.6	63.2	84.2	91.2	94.7	94.7	10.5
	FDAC1	31.3	50.0	70.8	79.2	85.4	89.6	31.3
	FELC0	38.9	66.7	79.6	85.2	85.2	90.7	14.8
	MSKT0	51.1	59.6	80.9	87.2	87.2	91.5	10.6
	MWBT0	44.0	56.0	84.0	92.0	96.0	98.0	26.0
	Overall	40.6	59.4	80.1	87.1	89.8	93.0	18.4
Bursts	FAKS0	55.9	86.4	98.3	98.3	98.3	98.3	10.2
	FDAC1	66.7	79.6	87.0	90.7	94.4	94.4	27.8
	FELC0	65.5	80.0	83.6	87.3	90.9	92.7	14.5
	MSKT0	58.0	76.0	88.0	90.0	94.0	96.0	12.0
	MWBT0	47.3	67.3	80.0	87.3	89.1	92.7	23.6
	Overall	58.8	78.3	87.9	91.2	93.8	95.2	17.6
Frication offsets	FAKS0	71.7	88.7	98.1	98.1	98.1	98.1	7.5
	FDAC1	47.4	65.8	76.3	86.8	86.8	94.7	23.7
	FELC0	56.8	65.9	72.7	75.0	77.3	84.1	15.9
	MSKT0	42.1	60.5	73.7	78.9	81.6	86.8	7.9
	MWBT0	43.5	58.7	73.9	76.1	84.8	87.0	19.6
	Overall	53.2	68.6	79.5	83.2	85.9	90.0	14.5
All types	FAKS0	55.0	79.3	93.5	95.9	97.0	97.0	9.5
	FDAC1	49.3	65.7	78.6	85.7	89.3	92.9	27.8
	FELC0	53.6	71.2	79.1	83.0	85.0	89.5	15.0
	MSKT0	51.1	65.9	81.5	85.9	88.1	91.9	10.4
	MWBT0	45.0	60.9	79.5	85.4	90.1	92.7	23.2
	Overall	50.9	69.0	82.8	87.4	90.1	92.9	17.0

window lengths for accurate detection. The overall detection rates at 10 ms are almost the same for all the speakers. Since we are going to use this method for extracting duration features for speaker recognition, its speaker-dependence may not affect the recognition performance; it might rather add additional speaker-dependent information.

Compared to LMD1, this method has improved detection rates in all levels of temporal accuracy for closure and frication offset landmarks. The number of insertion is also reduced from 59 to 48, resulting in 4% improvement. However, the detection rate of stop burst landmarks at 3 ms accuracy has reduced from 68% using LMD1 to 59% in this method. Jayan *et al* [46] have developed a method of stop landmark detection based on Gaussian mixture modeling of the speech spectrum, and evaluated it using the same set of TIMIT sentences as we used. Comparing with their results, the detection rates of stop burst landmarks with above 5 ms error and frication offset landmarks at 30 ms error are higher in Jayan's method by 3% and 6% respectively. However, the detection rates of stop closure landmarks in our method are higher by about 40%,

40% and 25% at the temporal error of 5, 10 and 15 ms respectively. The detection rates of frication offset landmarks at 5 and 10 ms also have improved from 19% and 40% in Jayan *et al*'s method, to 68% and 79% in out method.

Table 5.5 Number (percentage) of insertions in TIMIT sentences using LMD2

Insertion type	Number (percentage) of Insertions				Notes
	Closures	Bursts	Frication offsets	Total	
Pauses	13 (5.1)	13 (4.8)	11 (5.0)	37 (4.9)	Short pauses ('epi') mostly preceding nasals, and semivowels
Glottal stop	5 (2.0)	5 (1.8)	0 (0)	10 (1.3)	Brief closures denoted as 'q' in the TIMIT transcriptions
Nasals	8 (3.1)	8 (2.9)	6 (2.7)	22 (2.9)	Nasals ('n', 'm') closures and bursts, esp. those preceded or followed by pauses and semivowels
Fricatives	11 (4.3)	12 (4.4)	8 (3.6)	31 (4.1)	'dh', 'th' and 'v'
Semivowels and glides	4 (1.6)	4 (1.5)	2 (0.9)	10 (1.3)	'iy' followed by 'w', 'l' followed by 'iy'
Clicks	4 (1.6)	4 (1.5)	4 (1.8)	12 (1.6)	Clicks mostly in long pauses
Flap	2 (0.8)	2 (0.7)	1 (0.5)	5 (0.7)	'dx' sounds
Total	46 (18.4)	48 (17.6)	32 (14.5)	127 (17.0)	46 closure and 32 frication intervals, resulting in total insertion of 48 phonemes.

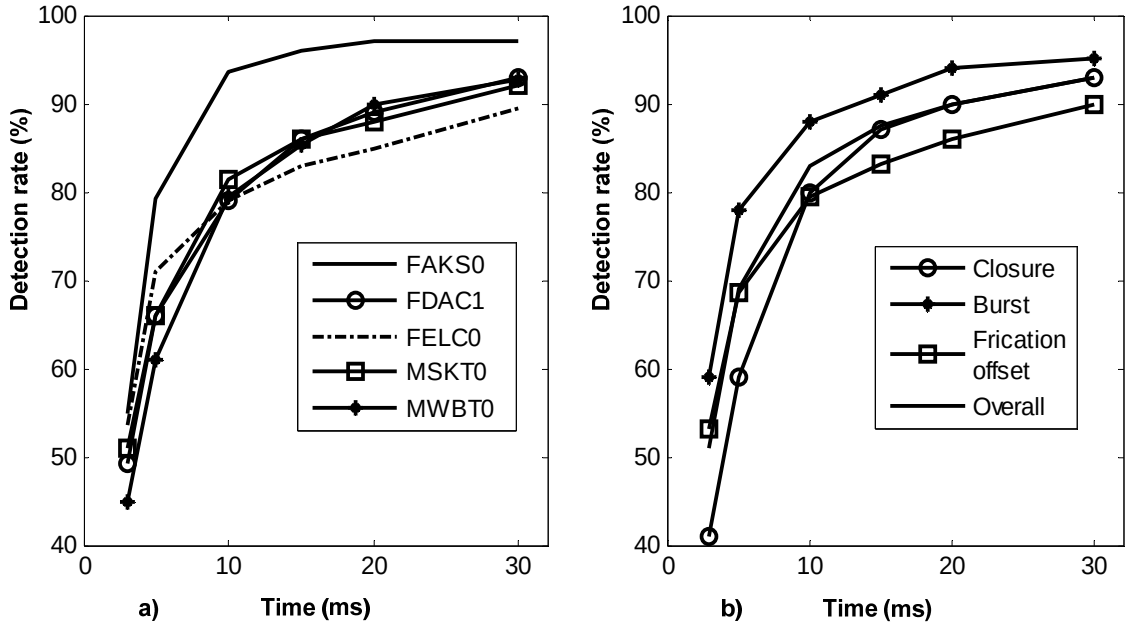

Fig. 5.13 Detection rates for TIMIT sentences using LMD2: a) speaker-wise rates, b) overall rates

5.9 Discussion

Spectral transitions in speech signal are manifested by rapid changes in signal characteristics such as intensity, periodicity and spectral envelope, and they represent major changes of articulation during the speech production process. Spectral rate-of-change functions derived from the magnitude spectrum, magnitude spectrum in 5 non-overlapping bands, and mel-filtered magnitude spectrum were studied for detecting spectral transitions. Median and mean smoothing along time axis of the original and mel-filtered magnitude spectrum were used to improve the detection process. Mel-filtering was investigated for smoothing harmonic transitions and other spectral discontinuities that occur along the frequency axis, while mean and median filtering were applied independently to reduce detection of false landmarks due to spectral transitions between pitch periods in voiced speech segments. A spectral slope parameter computed from the first difference of the filtered magnitude spectrum was used for detecting voicing onset and offset landmarks. The ROC of mel and median filtered squared magnitude spectrum, together with spectral slope and spectral flatness measure and log energy, was used for detecting stop consonant closure, burst and frication offset landmarks in clean VCV syllables and TIMIT sentences. Detection of stop landmarks on clean VCV utterances using the ROC of mel and median filtered squared magnitude spectrum and spectral slope resulted in overall detection rates of 53%, 75%, 91%, 95% and 97% within 3, 5, 10, 15 and 20 ms of the manual landmarks, respectively. Use of the ROC parameter, Wiener entropy and log energy for stop landmark detection in TIMIT sentences resulted in detection rates of 52%, 69%, 83%, 90%, and 93%, at 3, 5, 10, 20 and 30 ms accuracy respectively, and 18% insertion.

ROC function obtained directly from the first difference of the squared magnitude spectrum results in less distinct peaks at strong burst and voicing onset points, and it gives little indication of voicing offset instants. Median or mean smoothing of the squared magnitude spectrum along the time axis reduces noise and periodic spectral transitions in voiced speech, and ROC function so obtained gives improved indication of voicing onsets. Use of multiband ROC parameters helps to locate spectral transitions that occur in specific frequency bands; however, it is sensitive to narrow band noise and instantaneous pitch variations.

Filtering of the squared magnitude spectrum along the frequency axis using a triangular mel-filter bank emphasizes perceptually significant spectral transitions while masking harmonic structures. The advantage of this method is that its implementation is simple and it does not suffer from distortion or estimation errors as opposed to the methods that involve transformation or modelling of the speech spectrum. Besides, there is no temporal misalignment problem because the mel-filtering is done along the frequency axis. The stop landmark detection rates at temporal accuracies below 10 ms, obtained by picking peaks in the ROC of mel-filtered spectrum around closures, and voicing offsets and onsets, were promisingly high and the missed landmarks were

mainly due to missed closure segments or voicing onset/offsets landmarks. This suggests that the ROC parameter can accurately track abrupt spectral transients, and if we could come up with features for approximately locating the specific landmarks, a more precise location can be obtained from the ROC contour, and insertion rate can be reduced. One parameter we have found useful for locating stop landmarks is the product of log energy and Wiener entropy computed at high frame rate. It gives distinct dips during stop closure from which stop closures and bursts can be located with good temporal accuracy.

Chapter 6

SPEAKER MODELING AND RECOGNITION

6.1 Introduction

In Chapter 4, we have studied the variation of stop closure and burst durations across speakers using F-ratio analysis of variance tests, for both text-dependent and text-independent cases. The results indicate that there is a significant difference in mean closure and burst durations among speakers; hence these parameters could be potential candidates for speaker recognition. We have also presented in Chapter 5 a stop landmark detection method for extracting the stop duration parameters. In this chapter, we discuss a closed-set text-independent speaker recognition system using stop closure and burst durations in combination with MFCC features. As the ANOVA results show, duration features are highly text-dependent and they could give better performance in text-dependent recognition; however, our interest here is to improve the performance of a text-independent system, which generally has a lower performance compared to the text-dependent system.

Text-independent speaker recognition systems are often implemented using vector quantization (VQ) or Gaussian mixture models (GMM). VQ is preferred when small amount of training and testing data are used. However, GMM is superior in modeling the variations of unconstrained speech and is more robust to noise and mismatch conditions [12]. Reynolds and Rose [12] give two reasons for using GMM. First, the individual component densities of the GMM may represent some underlying set of acoustic classes such as vowels, nasals and fricatives. These classes reflect some general speaker-dependent vocal tract configurations that are useful for speaker discrimination. In our case, depending on the number of mixtures, the component GMMs are expected to model the duration parameters of the different stops (e.g., voiced, unvoiced, or individual stops). Secondly, any arbitrarily-shaped density function can be approximated by a linear combination of Gaussian basis functions.

Here, GMM is used for speaker recognition using stop closure and burst durations, and MFCC features. GMM is a stochastic model in which the distribution of the speaker's feature vectors is approximated by a linear combination of multi-dimensional Gaussian density functions, as described earlier in Section 2.4. The speaker model is represented by means, covariances, and weights of the Gaussian mixtures, estimated from the speaker's training feature vectors using expectation maximization algorithm. Speaker recognition is based on maximum log likelihood

score, computed using (2.19). Separate GMMs were generated for stop closure and burst durations, and MFCC features. The duration parameters were evaluated using manually obtained and automatically extracted duration features. MFCC features were extracted using the technique discussed in Section 3.3. Speaker recognition experiments were conducted on TIMIT data base. The following section discusses the training and testing stages, the database used, and details of the experiments carried out.

6.2 Training and Testing

In the training phase of speaker recognition, a set of speaker-dependent features are extracted from the training speech samples using suitable front-end processing techniques, and these features are used to generate speaker models. In our case, the features used are stop closure and burst durations, and MFCC features. The input speech was obtained at a sampling rate of 10 kHz, and it was pre-emphasized using a first-order high pass filter, with pre-emphasis coefficient of 0.97, to compensate for the -6 dB high frequency roll-off of voiced glottal sounds. Text-independent stop closure and burst durations were extracted from the resulting signal, using the landmark detector based on ROC of mel-filtered spectrum and closure detection, as described earlier in Section 5.7. Both the duration features are treated as one-dimensional vectors. To extract the MFCC features, low energy segments in the pre-emphasized speech signal were removed using the closure detector described in Section 5.7. Fourier transform of the filtered speech was computed using 512-pint DFT and, 20 ms Hanning window with 50% overlap. The magnitude spectrum was filtered using a bank of 40 triangular mel-filters with a minimum bandwidth of 133 Hz, and 13 cepstral coefficients were computed using (3.11). The standard 39-dimensional MFCC feature vector consisting of the 13 cepstral coefficients, 13 delta coefficients and 13 delta-delta coefficients was used for speaker recognition. Thus in our investigations, we have used three feature vectors: (i) closure duration as one-dimensional feature vector, (ii) burst duration as one-dimensional feature vector, and (iii) 39-dimemnsional MFCC feature vector. The numbers of observations in each set of feature vectors are generally different.

GMMs for each of the features were generated for each speaker using the expectation maximization algorithm described in Section 2.4. It may be noted that we use 1-dimensional GMM for the two duration features and 39-dimensional GMM for the MFCC feature vector. The initialization for estimating the GMM parameters was obtained using K-means clustering algorithm. In GMM fitting of a distribution, the number of components required to sufficiently approximate the distribution depends on the nature of the distribution. However, the number of maximum components in the model without resulting in computational errors depends on the number of observations, i.e. the amount of the training data [17]. During training, there were 30 – 70 observations for the two durations across speakers. It was found that use of two to six Gaussian

mixtures for modelling closure and burst durations gave best results. For MFCC feature vectors, there were 500 - 800 observations, and they were modelled using 8-component GMMs with diagonal covariance matrices.

During testing, stop closure and burst durations and MFCC features were extracted from the test utterances, and log likelihood of the test features were computed against each speaker model using (2.18). In order to compensate for speaker-independent score variability among speakers, the log likelihood scores were normalized using T-norm [10] as described in Section 2.5. For recognition using combination of duration and MFCC features, the final score was taken to be a weighted sum of the normalized scores from the three models, with weights of 0.75 for MFCC, 0.25 for the individual as well as combined duration features. The weights have been empirically assigned, based on the expected recognition performance of the features; with greater weight given to the features with lower probability of recognition error. Finally, speaker recognition decision was made based on maximum normalized log likelihood score.

Speaker recognition evaluation was carried out using TIMIT database. In this database, each person speaks 10 different sentences, out of which, 2 are common for all the speakers, 5 are common for only 7 speakers while the remaining three are different for each speaker. Initially, we tried a fully text-independent recognition in which the training utterances were different for some speakers. Testing using duration and MFCC features gave slight improvement over MFCC alone, but the performance of the duration parameters alone was poor. This may be attributed to the fact that duration features are highly text-dependent and features for one speaker extracted from a given set of utterances can overlap with the features for another speaker using a different set of utterances. For this reason, the duration parameters were evaluated for a set of speakers who have more utterances in common; so that the same set of training and testing utterances can be used for all the speakers. For this purpose, a set of 7 speakers who have 7 sentences in common were selected as reference speakers. A set of other 14 female and 16 male speakers were selected as impostors for T-norm score normalization. In order to study the performance of the features for different speakers, and training and test utterances, recognition experiments were conducted on five sets of reference speakers, with each set consisting of 7 different speakers. Recognition experiments using combination of MFCC and duration features were also conducted on a test set of 35 speakers, in order to investigate the contribution of stop closure and burst durations in a larger set of test speakers.

The performance of both the duration and MFCC features depends on the amount of training and testing data. Particularly, the duration parameters are one-dimensional and, a large amount of data would be required to extract enough features. To study this effect, two types of recognition experiments were conducted by varying the amount of training and testing data. In the first case, only the 7 sentences which are common to all the speakers in a given set were

considered, out of which 5 sentences were used for training and 2 for testing. In the second case, 3 sentences which are common to all speakers in a set were selected for testing and the remaining 7 sentences were used for training. Similar division of training and testing speech was used for the tests with 35 reference speakers. For testing, the duration features extracted from individual sentences were very short, and a concatenation of features from 2 or 3 sentences were used.

6.3 Results and disccusion

Recognition scores were obtained for different combinations of stop closure and burst durations, and MFCC features. Table 6.1 shows the recognition rates for each set of speakers obtained using 5 training sentences and 2 test sentences with T-norm. The total duration of the training speech was approximately 17 s, and that of the test speech was approximately 6 s. The number of mixtures column represents the number of GMM components used for modelling closure and burst durations. The number of mixtures used for MFCC features was 8 in all cases. Each speaker set consists of 7 speakers, and one recognition trial was conducted for each speaker. The numbers of female (F) and male (M) test speakers in each set are given in brackets in the table titles.

Table 6.1 Recognition rates (%) using detected durations with 17 s training and 6 s testing utterances, and T-norm (no. of test speakers: Set1 – 5 M, 2 F; Set2 – 5 M, 2 F; Set3 – 4 M, 3 F; Set4 – 4 M, 3 F; Set5 – 3 M, 4 F)

No. of mixtures for duration	Speaker set	Detection rate (%)						
		Closure	Burst	Closure + burst	MFCC	MFCC + closure	MFCC + burst	All three
1	Set1	29	14	14	100	100	100	100
	Set2	43	0	29	100	100	100	100
	Set3	29	0	43	100	100	100	100
	Set4	14	43	29	100	100	100	100
	Set5	29	0	14	57	86	57	57
	Overall	29	11	26	91	97	91	91
2	Set1	29	14	29	100	100	100	100
	Set2	14	14	43	100	100	100	100
	Set3	43	0	14	100	100	100	100
	Set4	14	14	29	100	100	100	100
	Set5	29	14	14	57	86	57	57
	Overall	26	11	26	91	97	91	91
3	Set1	29	14	29	100	100	100	100
	Set2	57	29	57	100	100	100	100
	Set3	29	29	14	100	100	100	100
	Set4	29	43	43	100	100	86	86
	Set5	14	0	0	57	71	57	57
	Overall	31	23	29	91	94	89	89

As results show, the performance of the duration features alone using short test utterances is very low, and this could be mainly due to lack of sufficient training and testing features. Closure duration gave relatively higher performance compared to frication duration, and no significant improvement was gained by using their combination. However, from the statistics of the individual recognition outcomes, it was observed in a number of tests that, when one feature misses, the other recognizes correctly, and vice versa, which suggests that the parameters may have complementary information. The recognition rates for MFCC alone, and duration and MFCC together were mostly similar. In speaker set5, which consisted of more number of female speakers, the recognition rate using MFCC alone was 52% and it was improved to 86% (2 more speakers identified correctly) using MFCC and closure duration features.

The recognition rates obtained using 7 training sentences (approximate duration of 25 s) are given in Table 6.2 and Table 6.3. The results in Table 6.2 were obtained using test features extracted from concatenation of 2 sentences. Table 6.3 shows results from recognition using 3 concatenated test sentences (duration 10 s). Compared to the results in Table 6.1, the recognition rates in this case are higher for both duration and MFCC features. Overall recognition rates of 43%, 34% and 34% were attained for closure and burst durations, and their combination respectively, using 2-component GMMs and approximately 10 s test utterances. The performances of the systems using either closure or burst duration with MFCC features were at least as good as that of the system using MFCC alone, while the system combining all the three features performed lower for some speakers.

The results show that stop closure and burst durations have speaker-dependent information. Their recognition rates with short test utterances are very low, but they can be combined with MFCC feature for improving speaker recognition. As the size of training and testing data increases, the performances of the both the duration and MFCC features improves. With respect to the number of GMM mixtures, closure duration was found to give better results with smaller number of components and its performance degraded significantly as the number of components exceeded 6, while the performance of burst duration was relatively less variable. There is a variation in the performances of the parameters across the 5 sets, which could be due to the difference in degree of similarity between speakers in each test set. Relatively, the results using closure duration are similar across the test sets. The overall low recognition rates of the duration features could be due to the following factors: (i) the amount of data used for training and testing was not enough to extract duration feature for speaker recognition, (ii) duration features are highly text-dependent and combining them in text-independent way can significantly reduce their speaker-discriminating capability, and (ii) the duration features might not have been accurately extracted.

Table 6.2 Recognition rates (%) using detected duration s with 25 s training and 6 s testing utterances, and T-norm (no. of test speakers: Set1 – 5 M, 2 F; Set2 – 5 M, 2 F; Set3 – 4 M, 3 F; Set4 – 4 M, 3 F; Set5 – 3 M, 4 F)

No. of mixtures For duration	Speaker set	Detection rate (%)						
		Closure	Burst	Closure + burst	MFCC	MFCC + closure	MFCC + burst	All three
2	Set1	57	0	43	100	100	100	100
	Set2	14	14	29	100	100	100	100
	Set3	57	14	14	100	100	100	100
	Set4	43	57	43	100	100	100	100
	Set5	14	43	43	86	86	100	86
	Overall	37	26	34	97	97	100	97
3	Set1	29	14	29	100	100	100	100
	Set2	29	29	43	100	100	100	100
	Set3	29	14	29	100	100	100	100
	Set4	29	29	29	100	100	100	100
	Set5	43	43	29	86	86	100	71
	Overall	31	26	31	97	97	100	94

Table 6.3 Recognition rates (%) using detected durations with 25 s training and 10 s testing utterances, and T-norm (no. of test speakers: Set1 – 5 M, 2 F; Set2 – 5 M, 2 F; Set3 – 4 M, 3 F; Set4 – 4 M, 3 F; Set5 – 3 M, 4 F)

No. of mixtures for durations	Speaker set	Detection rate (%)						
		Closure	Burst	Closure + burst	MFCC	MFCC + closure	MFCC + burst	All three
2	Set1	43	0	14	100	100	100	100
	Set2	43	57	57	100	100	100	100
	Set3	43	14	14	100	100	100	100
	Set4	57	43	57	100	100	100	100
	Set5	29	57	29	86	86	100	86
	Overall	43	34	34	97	97	100	97
3	Set1	29	0	14	100	100	100	100
	Set2	29	29	43	100	100	100	100
	Set3	29	14	29	100	100	100	100
	Set4	29	29	29	100	100	100	100
	Set5	14	14	14	86	86	100	86
	Overall	26	17	26	97	97	100	97
4	Set1	43	0	43	100	100	100	100
	Set2	14	0	0	100	100	100	100
	Set3	43	14	71	100	100	100	100
	Set4	0	29	0	100	100	100	86
	Set5	14	0	14	86	86	86	57
	Overall	23	9	26	97	97	97	89

Table 6.4 Recognition rates (%) using manual durations with 25 s training and 10 s testing utterances, and T-norm (no. of test speakers: Set1 – 5 M, 2 F; Set2 – 5 M, 2 F; Set3 – 4 M, 3 F; Set4 – 4 M, 3 F; Set5 – 3 M, 4 F)

No. of mixtures for duration	Speaker set	Detection rate (%)						
		Closure	Burst	Closure + burst	MFCC	MFCC + closure	MFCC + burst	All three
2	Set1	43	43	71	100	100	100	100
	Set2	57	29	43	100	100	100	100
	Set3	57	29	43	100	100	100	100
	Set4	29	29	43	100	100	100	100
	Set5	43	43	43	86	86	100	100
	Overall	46	34	49	97	97	100	100
3	Set1	43	43	71	100	100	100	100
	Set2	57	29	43	100	100	100	100
	Set3	57	29	43	100	100	100	100
	Set4	29	29	43	100	100	100	100
	Set5	43	43	43	86	86	100	86
	Overall	46	34	49	97	97	100	97
4	Set1	57	29	71	100	100	100	100
	Set2	71	43	43	100	100	100	100
	Set3	29	14	43	100	100	100	100
	Set4	0	29	14	100	100	100	100
	Set5	57	57	57	86	86	100	86
	Overall	43	34	46	97	97	100	97
6	Set1	29	29	43	100	100	100	100
	Set2	57	29	57	100	100	100	100
	Set3	43	29	43	100	100	100	100
	Set4	14	0	14	100	100	100	86
	Set5	71	71	86	86	86	100	86
	Overall	43	31	49	97	97	100	94

In order to examine if the problem was due to temporal inaccuracies in landmark detection, recognition tests were conducted using duration features extracted manually from the TIMIT transcriptions. Table 6.4 and Table 6.5 show detection rates obtained using approximately 25 s training and 10 s data, with and without T-norm score normalization respectively. Cumulatively, the performances in this case are slightly higher for all the features as compared to the results obtained using automatically detected duration features. Recognition rates of 45%, 37% and 51% were obtained for closure, burst and their combination respectively, using 2-component GMMs and without test normalization. In some cases, recognition rates of up to 85%

were attained using closure and burst duration features together. The normalized recognition scores obtained using combination of MFCC and duration, were at least the same as those obtained using only MFCC features. It is to be noticed that there were situations in which speakers missed by the MFCC features were consistently recognized either by closure or burst duration. Table 6.6 shows recognition results obtained using training data of approximately 17 s and test data of about 6 s. The rates due to the duration features alone were still low; however, the scores obtained using combination of MFCC and duration are mostly better than the scores obtained using only MFCC features.

Table 6.5 Recognition rates (%) using manual duration with 25 s training and 10 s testing utterances, and without T-norm (no. of test speakers: Set1 – 5 M, 2 F; Set2 – 5 M, 2 F; Set3 – 4 M, 3 F; Set4 – 4 M, 3 F; Set5 – 3 M, 4 F)

No. of mixtures for durations	Speaker set	Detection rate						
		Closure	Burst	Closure + burst	MFCC	MFCC + closure	MFCC + burst	All three
2	Set1	57	57	86	100	100	100	100
	Set2	57	29	43	100	100	100	100
	Set3	43	29	43	100	100	100	100
	Set4	29	29	43	100	100	100	100
	Set5	43	43	43	86	86	100	86
	Overall	46	37	51	97	97	100	97
3	Set1	57	57	86	100	100	100	100
	Set2	57	29	43	100	100	100	100
	Set3	43	29	43	100	100	100	100
	Set4	29	29	43	100	100	100	86
	Set5	43	43	43	86	86	100	86
	Overall	46	37	51	97	97	100	94
4	Set1	57	29	71	100	100	100	100
	Set2	71	43	43	100	100	100	100
	Set3	29	14	43	100	100	100	100
	Set4	0	29	14	100	100	100	71
	Set5	57	71	71	86	86	100	86
	Overall	43	37	49	97	97	100	91
6	Set1	29	29	43	100	100	100	100
	Set2	57	29	57	100	100	100	100
	Set3	43	29	43	100	100	100	100
	Set4	14	0	14	100	86	100	86
	Set5	71	71	86	86	86	100	100
	Overall	43	31	49	97	94	100	97

Table 6.6 Recognition rates (%) using manual durations with 17 s training and 6 s testing utterances, and T-norm (no. of test speakers: Set1 – 5 M, 2 F; Set2 – 5 M, 2 F; Set3 – 4 M, 3 F; Set4 – 4 M, 3 F; Set5 – 3 M, 4 F)

No. of mixtures for durations	Speaker set	Detection rate						
		Closure	Burst	Closure + burst	MFCC	MFCC + closure	MFCC + burst	All three
3	Set1	43	71	57	100	100	100	100
	Set2	14	14	0	100	100	100	86
	Set3	43	14	29	100	100	100	100
	Set4	14	29	29	71	86	71	86
	Set5	43	0	14	86	86	100	100
	Overall	31	26	26	91	94	94	94
4	Set1	29	29	29	100	100	100	100
	Set2	29	29	43	100	100	100	86
	Set3	43	14	14	100	100	100	100
	Set4	14	14	29	71	86	71	86
	Set5	14	0	14	86	86	86	100
	Overall	26	17	26	91	94	91	94
4	Set1	29	29	29	100	100	100	100
	Set2	14	14	43	100	100	100	86
	Set3	14	29	29	100	100	100	100
	Set4	14	14	14	71	100	71	57
	Set5	29	29	43	86	71	100	71
	Overall	20	23	31	91	94	94	83

Comparing the results in Table 6.4 and Table 6.5, it can be observed that the recognition rates for MFCC are improved with T-norm, while the rates for the duration features are higher without T-norm. The T-norm could have the effect of normalizing out the speaker-dependent duration variations across different utterances. Improved recognition rates could be obtained by applying the T-norm only to the MFCC scores; however, combining the normalized MFCC scores with the un-normalized duration scores may not be a simple task.

In order to find out the contribution of the duration features on a larger set of test speakers, text-independent recognition experiments using combination of MFCC and duration features were carried out on the entire set of 35 speakers. Two recognition experiments were conducted using different amount of training and testing speech. In first case, 8 out of the 10 sentences (25 s duration) were used for training and 2 sentences (6 s duration) were used for testing. In the second case, 7 sentences (20 s duration) were used for training and 3 sentences (10 s duration) for testing. The duration features were modelled using 2-component GMMs, and MFCC features were modelled using 8 Gaussian mixtures as before. In both cases, 2 recognition trials were conducted for each speaker, resulting in a total of 70 trials in each experiment, and the log likelihood scores were computed without T-norm. Both manually measured and automatically

extracted duration features were used. The results are summarized in Table 6.7. The performances of MFCC combined with stop duration features are mostly higher than, and at least the same as, that of MFCC features. Using combination of MFCC and automatically extracted closure duration resulted in an improvement of 3% over the MFCC features alone. The performance of all the features has improved with increasing the amount training speech. The results show that the improvement obtained using automatically extracted duration features are higher than that of the manually measured durations. This may be due to the fact that the temporal errors in the landmark detection are speaker-dependent, and they can add speaker information to the duration features which can improve recognition performance.

Table 6.7 Recognition results (%) using a test set of 35 speakers, without T-norm

Extraction method	Length of training utterance (s)	Length of Test utterance (s)	Detection rate			
			MFCC	MFCC + closure	MFCC + burst	All three
Manual	25	6	93	94	93	94
	20	10	90	90	90	90
Detected	25	6	93	96	93	94
	20	10	90	94	90	93

6.4 Summary

The recognition results obtained using automatically detected and manually extracted duration features show that stop duration parameters have speaker-dependent information. The performances of MFCC and duration features together were generally better than that of the MFCC features alone, especially with short training and test utterances. This shows that the duration features convey speaker information which is complementary to that of MFCC features. Besides, the duration features are one-dimensional and training and testing takes a relatively small time; therefore, they can be used together with the MFCC features to improve speaker recognition without adding significant effect on the system complexity. In the experiments with the same training and testing utterances, the recognition rates obtained using automatically detected closure and burst durations were comparable to the results obtained using manually extracted features. This indicates that the landmark detector was able to detect stop landmarks with acceptable temporal accuracy, and small inaccuracy may not have significant effect on performance of the speaker recognition system. The results obtained using 35 test speakers also indicated the temporal errors during landmark detection may be speaker-dependent and they can contribute to the improvement of speaker recognition using duration features.

Chapter 7

SUMMARY AND CONCLUSION

The objective of this project was to investigate the variation of stop duration parameters using automatic landmark detection for improving speaker recognition. For this purpose, the variation of stop closure durations across speakers was studied using analysis of variance (ANOVA) tests on VCV syllables and sentences recorded at different speaking rates. The effect of context variation on stop closure and burst durations was analyzed using two-way ANOVA on VCV utterances and TIMIT sentences. In order to extract the duration features, a landmark detection technique based on the rate-of-change of mel-filtered spectrum, spectral flatness measure, spectral slope and signal energy was developed. The method was evaluated on VCV syllables and TIMIT sentences. Finally, speaker recognition experiments were carried out on TIMIT database using Gaussian mixture modeling of stop closure and burst durations, and MFCC features.

The one-way ANOVA results showed that text-dependent stop closure duration varies significantly across speakers. However, the results also showed that the speaking rate and context variability have significant effect on closure and burst durations. The variation of stop closure duration was found to be high among different speakers compared to the variations due to vowel and stop types, while burst duration was found to be highly dependent on the type of stop. The results suggest that stop duration parameters convey speaker-dependent information and could be used in combination with the standard features for improving speaker recognition.

Two types of landmark detectors were developed for detecting stop closure, burst and burst offset: 1) based one ROC of mel-filtered squared magnitude spectrum and voicing detection, and 2) based on ROC of mel-filtered squared magnitude spectrum and closure detection. In both cases, Wiener entropy (a measure of spectral flatness) of the magnitude spectrum was used for validating stop landmarks. A measure of spectral slope computed from the first difference of the mel-filtered magnitude spectrum along the frequency axis was used for detecting voicing onset/offset landmarks. Closure segments were detected using a new parameter computed as the product of Wiener entropy and log energy. Detection of stop landmarks in VCV syllables using ROC and voicing detection resulted in overall detection rates of 53%, 75%, 90%, 95% and 97% respectively within 3, 5, 10, 15 and 30 ms of the manually labeled landmarks. The detection rates on TIMIT sentences using the same method were 52%, 68%, 79%, 84%, 87% and 91% at temporal accuracies of 3, 5, 10, 15, 20 and 30 ms respectively. Using ROC and closure detection,

stop landmarks in TIMIT sentences were detected at rates of 52%, 69%, 83%, 87%, 90% and 93% within 3, 5, 10, 15, 20 and 30 ms of the manual landmarks, respectively. The results show that the ROC of mel-filtered magnitude spectrum is effective in tracking abrupt acoustic transitions. Mel-filtering of the squared magnitude spectrum along the frequency axis improves landmark detection by enhancing the perceptually important spectral transitions and, smoothing harmonic structure and noise.

The speaker recognition experiments were carried out on TIMIT database using Gaussian mixture modeling of MFCC features and automatically detected and manually extracted stop closure and burst duration features. The number of speakers in a test set was 7, and 5 speaker sets were taken in order to investigate the performances of the features for different speaker sets, and training and test utterances. The performance of duration features alone was not satisfactory. However, compared to the rates due to MFCC features alone, an improvement of 4% was obtained by combing MFCC with closure durations. This shows that stop closure and burst durations have speaker information which may not be conveyed by the MFCC parameters; therefore, they can be used for improving speaker recognition. However, the performances of the closure and burst durations were very low, which indicates that such duration parameters alone could not be useful for text-independent speaker recognition.

This work has focused on the detection of stop landmarks for extracting stop closure and burst durations for text-independent speaker recognition, using short training and test speech. However, the project can be extended to improve the performance of the landmark detection method and to study additional duration features for speaker recognition. The landmark detector can be further investigated for detecting different types of speech landmarks by incorporating additional phonetic features. The duration features can be studied using a larger amount of training and testing data for both text-dependent and text-independent speaker recognition. Moreover, improved recognition performance may be attained by devising a better scheme for combining the scores obtained using different recognition parameters.

Appendix A

RESULTS FROM ANOVA OF STOP CLOSURE AND BURST DURATIONS

A.1 One-Way ANOVA of Closure Duration

Table A.1 Means and square root of variances of stop closure duration at normal speaking rate (closure duration in ms, Sp = speaker)

Utter-ance	Stop closure	Mean (ms)					$\sqrt{\text{variance}}$ (ms)				
		Sp1	Sp2	Sp3	Sp4	Sp5	Sp1	Sp2	Sp3	Sp4	Sp5
VCV01	/aba/	118.7	86.7	117.4	80.4	89.3	5.1	5.1	3.6	8.5	8.2
	/ada/	107.3	84.9	115.9	74.4	87.3	6.1	7.5	9.1	4.8	6.3
	/aka/	109.9	95.7	112.1	81.4	87.9	9.3	6.4	5.8	7.7	7.3
	/apa/	118.0	119.0	121.1	100.9	94.1	5.0	9.1	5.8	5.5	5.2
	/ata/	119.1	118.9	116.9	106.9	86.6	4.4	8.8	6.2	6.3	7.6
VCV02	/ibi/	111.1	103.7	86.3	80.4	98.1	5.7	9.4	4.8	10.0	8.9
	/idi/	112.6	89.9	89.0	69.0	100.9	9.1	4.0	7.0	5.8	8.8
	/iki/	104.9	88.0	89.0	87.3	90.0	7.3	8.1	6.7	4.6	5.7
	/ipi/	117.7	120.9	99.9	96.0	106.3	7.2	6.2	8.2	5.3	7.9
	/iti/	116.4	97.1	107.0	81.6	87.7	5.3	9.8	4.7	12.5	10.2
S01	/edi-/	50.3	25.1	52.7	42.3	36.3	5.5	6.8	5.8	8.4	1.9
	/-ito-|	57.1	38.6	44.1	54.3	42.4	5.6	3.5	11.4	3.0	5.6
	/-oppe-|	68.9	59.4	54.0	57.7	64.8	7.6	13.1	6.2	3.7	5.0
	/-ed|	68.6	48.3	53.4	35.8	31.7	11.0	4.2	10.4	9.0	10.4
	/-ape-/	94.3	81.3	72.1	63.7	77.6	4.3	5.2	8.8	5.1	6.2
S02	/-api-/	79.6	60.0	75.9	52.3	63.9	4.5	6.7	8.3	4.3	5.3
	/-ita-/	44.9	29.0	35.0	35.6	-	6.7	2.2	7.8	6.9	-
	/-ity/	49.9	22.0	52.3	49.0	36.3	7.8	1.6	7.3	5.1	2.5
	/-opu-/	90.0	62.3	85.3	53.6	69.7	3.4	3.1	4.1	8.6	8.0
	/-ate-/	58.6	35.9	59.7	43.7	51.3	5.1	5.4	11.3	3.0	10.0
S03	/-ick-/	44.0	49.9	58.9	35.0	44.4	5.0	4.9	15.3	5.5	3.4
	/-ed/	51.4	33.3	37.9	47.4	-	6.1	4.5	11.4	2.3	-
	/-up/	86.3	71.3	72.3	55.7	57.6	14.9	6.7	8.0	6.4	4.8
	/-ack/	90.7	75.7	73.6	52.7	60.0	7.5	6.3	2.1	1.8	4.5
	/-age/	71.0	52.6	59.1	47.4	58.7	3.6	8.1	10.9	5.5	8.8
S04	/ab-/	103.9	90.3	67.1	57.4	83.1	6.8	6.4	6.2	4.1	4.8
	/-ed-/	60.1	77.0	-	37.7	-	5.1	4.3	-	3.1	-
	/-aby-/	63.7	65.1	53.6	54.1	58.6	2.8	4.2	7.1	7.8	5.4

Utter-ance	Stop closure	Mean (ms)					$\sqrt{\text{variance}}$ (ms)				
		Sp1	Sp2	Sp3	Sp4	Sp5	Sp1	Sp2	Sp3	Sp4	Sp5
S05	/-oto-/	39.7	22.9	35.3	44.7	27.0	1.1	3.1	4.5	7.2	3.5
	/-oco-/	66.4	50.9	60.4	39.3	45.6	8.0	9.5	5.5	2.8	4.4
	/-opi-/	100.3	56.7	65.7	61.6	70.4	9.8	5.7	12.8	6.0	6.5
	/-ied/	82.0	111.4	55.6	72.1	78.3	10.0	6.5	9.0	5.9	10.2
	/-ook/	110.4	100.1	102.3	90.1	88.4	12.0	7.5	4.8	9.2	7.3
S06	/abi-/	105.1	91.7	87.1	52.0	79.6	3.9	9.2	10.9	3.6	3.1
	/-ide-/	55.4	37.9	31.6	46.7	43.7	5.3	6.1	7.9	4.5	4.6
	/edu-/	70.9	51.1	60.3	48.7	44.1	2.8	2.8	7.8	8.3	4.2
	/-uca-/	51.3	47.1	70.7	49.2	59.7	6.5	3.6	8.6	5.2	6.2
	/-ate-/	61.7	26.9	48.9	49.7	48.7	7.1	13.0	4.4	3.9	5.7

Table A.2 ANOVA of stop closure duration at normal speaking rate (SS_W = sum of squares within speakers, SS_B = sum of squares between speakers, MSS_W = mean of squares within speakers, MSS_B = mean of squares between speakers, $\alpha = 0.05$ for all cases, duration values in ms)

Utterance	Stop closure	SS_W	MSS_W	SS_B	MSS_B	F-ratio	P-value	$F_{critical}$
VCV01	/aba|	1227.7	40.9	9221.0	2305.2	56.3	1.6e-13	2.7
	/ada|	1436.3	47.8	8161.6	2040.4	42.6	5.8e-12	2.7
	/aka|	1635.7	54.5	5050.6	1262.6	23.2	8.3e-09	2.7
	/apa|	1190.6	39.6	4215.6	1053.9	26.5	1.8e-09	2.7
	/ata|	1403.1	46.7	5370.7	1342.6	28.7	7.2e-10	2.7
VCV02	/ibi|	1932.3	64.4	4411.6	1102.9	17.1	2.1e-07	2.7
	/idi|	1555.4	51.8	7307.2	1826.8	35.2	6.2e-11	2.7
	/iki|	1306.3	43.5	1514.6	378.6	8.7	8.7e-05	2.7
	/ipi|	1486.5	49.5	3309.7	827.4	16.7	2.7e-07	2.7
	/iti|	2439.7	81.3	5579.2	1394.8	17.2	2.0e-07	2.7
S01	/edi-|	1100.5	36.6	3487.3	871.8	23.8	6.2e-09	2.7
	/-ito-|	1282.5	42.7	1788.9	447.2	10.46	2.0e-05	2.69
	/-oppe-|	1820.8	62.7	955.4	238.8	3.80	1.3e-02	2.70
	-ed|	1969.6	93.7	4508.3	1127.1	12.02	3.1e-05	2.84
	/-ape-|	883.0	35.3	2964.8	741.2	20.99	1.1e-07	2.76
S02	/-api-|	1090.9	36.4	3566.7	891.7	24.5	4.4e-09	2.7
	/-ita-/	952.6	41.4	900.1	225.0	5.4	3.1e-03	2.8
	/-ity/	889.7	31.8	3705.0	926.3	29.2	1.3e-09	2.7
	/-opu-/	1056.0	35.2	6577.0	1644.2	46.7	1.8e-12	2.7
	/-ate-/	1350.1	51.9	2853.0	713.3	13.7	3.8e-06	2.7
S03	/-ed/	441.4	49.1	5546.9	1386.7	28.3	4.1e-05	3.6
	/-ed/	1163.1	50.6	1473.2	368.3	7.28	6.1e-04	2.80
	/-up/	2369.4	78.9	4348.7	1087.1	13.77	1.7e-06	2.69
	/-ack/	738.0	24.6	6102.6	1525.6	62.02	4.5e-14	2.69
	/-age/	1779.7	68.4	1392.1	348.0	5.08	3.7e-03	2.74
S04	/ab-/	995.7	33.1	9512.4	2378.1	71.65	6.5e-15	2.69

Utterance	Stop closure	SS_W	MSS_W	SS_B	MSS_B	F-ratio	P-value	$F_{critical}$
S04	/-ed-/	324.2	20.2	5438.0	1359.5	67.08	8.6e-10	3.01
	/-aby-/	996.5	33.2	792.4	198.1	5.96	1.2e-03	2.69
S05	/-oto-/	565.1	18.8	2255.6	563.9	29.93	4.4e-10	2.69
	/-oco-/	1265.3	42.1	3375.3	843.8	20.00	4.1e-08	2.69
	/-opi-/	2213.7	73.7	8252.1	2063.0	27.96	9.7e-10	2.69
	/-ied/	1771.7	68.1	11551.6	2887.9	42.38	5.0e-11	2.74
	/-ook/	2174.5	72.4	2312.5	578.1	7.98	1.7e-04	2.69
S06	/abi-/	1448.8	48.3	10892.6	2723.1	56.39	1.6e-13	2.69
	/-ide-/	1015.1	33.8	2280.7	570.1	16.85	2.4e-07	2.69
	/edu-/	987.4	32.9	3161.5	790.3	24.01	5.5e-09	2.69
	/-uca-/	1143.9	39.4	2595.5	648.9	16.45	3.9e-07	2.70
	/-ate-/	1703.9	58.7	4442.8	1110.7	18.90	9.6e-08	2.70

A.2 Two-Way ANOVA of Stop Closure Durations at Different Speaking Rates

Table A.3 Means and square root of variances of stop closure duration at different speaking rates (Sp = speaker, var = variance)

Utter-ance	Stop closure	Speaking Rate	Mean (ms)				$\sqrt{var}$ (ms)			
			Sp1	Sp2	Sp3	Total	Sp1	Sp2	SP3	Overall
S11	/edi-/	Slow	34.8	83.2	71.0	63.0	3.3	1.8	9.3	21.9
		Normal	36.2	43.8	50.2	43.4	1.9	5.2	4.3	7.0
		Fast	27.2	34.2	22.2	27.9	3.9	2.5	6.9	6.8
		Overall	32.7	53.7	47.8	44.8	5.0	22.2	21.7	19.9
	/-oto-/	Slow	62.8	112.2	57.8	77.6	6.3	8.7	3.0	26.1
		Normal	39.4	64.8	36.4	46.8	3.9	1.8	3.4	13.5
		Fast	28.0	40.4	32.4	33.6	4.5	3.3	6.7	7.1
		Overall	43.4	72.5	42.2	52.7	15.7	31.3	12.3	25.3
	/-oco-/	Slow	76.0	124.4	84.2	94.8	3.7	15.0	8.5	23.8
		Normal	59.4	56.6	61.8	59.2	3.0	4.8	2.5	4.0
		Fast	40.0	48.0	53.0	47.0	3.4	6.3	5.4	7.3
		Overall	58.5	76.3	66.3	67.0	15.5	36.5	14.7	25.0
	/-opie-/	Slow	151.8	145.4	138.6	145.3	10.4	16.5	6.1	12.3
S11	/-opie-/	Normal	73.6	89.4	92.6	85.2	7.5	7.7	9.6	11.5
		Fast	49.6	63.2	65.6	59.4	4.3	2.6	11.7	10.0
		Overall	91.7	99.3	98.9	96.6	45.7	36.8	32.4	38.0
	/-ook/	Slow	122.4	231.0	132.4	161.9	8.6	56.2	8.6	59.3
		Normal	89.4	110.8	92.8	97.6	3.6	4.4	4.0	10.4
		Fast	87.4	91.8	81.4	86.8	9.7	7.6	4.4	8.3
		Overall	99.7	144.5	102.2	115.5	18.1	70.7	23.3	49.0
S12	/-api-/	Slow	158.2	169.4	140.0	155.9	9.4	27.6	11.6	20.9
		Normal	77.0	73.0	73.4	74.5	5.6	2.7	3.0	4.1

Utter-ance	Stop closure	Speaking Rate	Mean (ms)				$\sqrt{\text{var}}$ (ms)			
			Sp1	Sp2	Sp3	Total	Sp1	Sp2	SP3	Overall
S12	/-api-/	Fast	45.2	53.0	66.8	55.0	1.9	5.2	2.9	9.8
		Overall	93.5	98.5	93.4	95.1	49.6	54.7	34.8	46.1
	/-ita-/	Slow	53.4	213.0	62.8	109.7	3.8	29.9	4.3	77.4
		Normal	39.8	42.2	24.6	35.5	5.9	6.4	7.5	10.1
		Fast	24.4	31.0	26.6	27.3	3.2	3.9	10.1	6.7
		Overall	39.2	95.4	38.0	57.5	12.9	87.8	19.5	58.0
	|-ity|	Slow	63.8	255.2	52.8	123.9	6.2	16.4	5.1	96.7
		Normal	40.8	74.4	45.8	53.7	3.6	6.3	5.4	16.1
		Fast	27.4	54.8	42.6	41.6	3.6	4.9	3.4	12.2
		Overall	44.0	128.1	47.1	73.1	16.1	93.9	6.2	66.7
	/-opu-/	Slow	107.4	107.8	122.8	112.7	8.3	21.3	6.3	14.7
		Normal	80.0	60.6	90.4	77.0	4.7	3.5	6.8	13.7
		Fast	68.6	52.0	72.2	64.3	5.9	2.2	10.1	11.1
		Overall	85.3	73.5	95.1	84.6	17.9	27.9	22.9	24.4
	/-ate-/	Slow	51.6	142.2	54.6	82.8	3.6	22.1	3.3	45.1
		Normal	42.4	51.2	45.8	46.5	4.6	5.8	12.7	8.7
		Fast	39.6	39.4	52.8	43.9	2.5	4.0	4.6	7.4
		Overall	44.5	77.6	51.1	57.7	6.3	49.1	8.4	31.8
S03	/aba-/	Slow	191.0	91.0	117.6	133.2	8.2	5.8	8.6	44.3
		Normal	79.4	69.0	94.4	80.9	5.7	2.7	3.8	11.5
		Fast	59.4	51.4	79.4	63.4	3.0	3.4	2.1	12.5
		Overall	109.9	70.5	97.1	92.5	60.2	17.2	17.1	40.2
	/aba-/	Slow	75.2	86.4	58.0	73.2	3.6	8.0	4.5	13.2
		Normal	72.0	59.8	52.2	61.3	4.1	3.3	7.0	9.7
		Fast	55.4	50.8	49.6	51.9	4.7	3.0	3.9	4.5
		Overall	67.5	65.7	53.3	62.2	9.8	16.4	6.1	13.0
	/ucu/	Slow	83.6	160.0	59.0	100.9	4.0	14.9	5.3	45.4
		Normal	52.4	95.8	56.8	68.3	4.8	4.3	1.3	20.5
		Fast	43.8	62.4	38.2	48.1	4.3	3.3	2.4	11.2
		Overall	59.9	106.1	51.3	72.4	18.2	42.8	10.2	36.2
S14	/-icke-/	Slow	81.6	146.8	52.0	93.5	4.7	13.2	6.6	41.8
		Normal	39.6	57.8	44.2	47.2	4.0	6.6	4.6	9.3
		Fast	31.6	40.6	37.8	36.7	3.2	7.2	5.0	6.3
		Overall	50.9	81.7	44.7	59.1	23.0	49.0	7.9	34.9
	/up/	Slow	122.2	150.8	122.4	131.8	3.0	8.5	11.8	16.0
		Normal	62.8	77.2	57.4	65.8	9.7	2.3	4.3	10.4
		Fast	45.2	48.2	40.2	44.5	1.9	4.9	2.4	4.6
		Overall	76.7	92.1	73.3	80.7	34.5	45.0	37.3	39.2
	/-acka-/	Slow	119.8	144.0	129.8	131.2	8.7	15.2	5.7	14.2
		Normal	65.4	78.0	85.4	76.3	5.9	3.4	4.3	9.6
		Fast	46.0	51.0	54.4	50.5	3.0	3.4	5.4	5.2
		Overall	77.1	91.0	89.9	86.0	32.9	41.3	32.4	35.5

Table A.4 Two-way ANOVA of stop closure duration at different speaking rates (Sp = speaker, α = 0.05)

Utterance	Stop closure	Source of variation	Sum of squares (ms)	Degree of freedom	Mean of squares (ms)	F-ratio	P-value	$F_{critical}$
S11	/edi-/	Rate	9298.9	2	4649.4	190.9	6.8e-20	3.3
		Speakers	3516.0	2	1758.0	72.2	2.5e-13	3.3
		Interaction	3674.8	4	918.7	37.7	2.1e-12	2.6
		Within	876.4	36	24.3	-	-	-
	/-oto-/	Rate	15282.7	2	7641.3	299.6	3.6e-23	3.3
		Speakers	8811.9	2	4405.9	172.7	3.5e-19	3.3
		Interaction	3059.0	4	764.7	29.9	5.1e-11	2.6
		Within	918.0	36	25.5	-	-	-
	/-oco-/	Rate	18545.2	2	9272.6	194.6	4.9e-20	3.3
		Speakers	2405.5	2	1202.7	25.2	1.4e-07	3.3
		Interaction	4801.9	4	1200.4	25.2	5.2e-10	2.6
		Within	1715.2	36	47.6	-	-	-
	/-opie-/	Rate	58159.2	2	29079.6	331.9	6.4e-24	3.3
		Speakers	558.7	2	279.3	3.2	5.3e-02	3.3
		Interaction	1656.3	4	414.0	4.7	6.1e-03	2.6
		Within	3154.0	36	87.6	-	-	-
	/-ook/	Rate	49409.2	2	24704.6	63.1	1.7e-12	3.3
		Speakers	19026.1	2	9513.0	24.3	2.1e-07	3.3
		Interaction	18595.4	4	4648.8	11.8	2.9e-06	2.6
		Within	14098.4	36	391.6	-	-	-
S12	/-api-/	Rate	85895.0	2	42947.5	361.0	1.5e-24	3.3
		Speakers	253.4	2	126.7	1.1	0.4	3.3
		Interaction	3193.3	4	798.3	6.7	3.8e-04	2.6
		Within	4282.8	36	119.0	-	-	-
	/-ita-/	Rate	61813.2	2	30906.6	234.3	2.3e-21	3.3
		Speakers	32273.2	2	16136.6	122.3	8.9e-17	3.3
		Interaction	48951.6	4	12237.9	92.8	1.9e-18	2.6
		Within	4749.2	36	131.9	-	-	-
	/-ity/	Rate	59308.9	2	29654.5	578.2	4.4e-28	3.3
		Speakers	68298.1	2	34149.1	665.8	3.7e-29	3.3
	/-ity/	Interaction	66407.3	4	16601.8	323.7	1.1e-27	2.6
		Within	1846.4	36	51.3	-	-	-
	/-opu-/	Rate	18884.0	2	9442.0	108.4	5.8e-16	3.3
		Speakers	3531.5	2	1765.8	20.3	1.3e-06	3.3
		Interaction	687.6	4	171.9	2.0	0.1	2.6
		Within	3135.2	36	87.1	-	-	-
	/-ate-/	Rate	14185.7	2	7092.9	83.0	3.3e-14	3.3
		Speakers	9200.5	2	4600.3	53.9	1.5e-11	3.3
		Interaction	18071.3	4	4517.8	52.9	1.4e-14	2.6
		Within	3075.2	36	85.4	-	-	-

Utterance	Stop closure	Source of variation	Sum of squares (ms)	Degree of freedom	Mean of squares (ms)	F-ratio	P-value	$F_{critical}$
S13	/aba-/	Rate	39556.3	2	19778.2	702.7	1.4e-29	3.3
		Speakers	12162.8	2	6081.4	216.1	8.8e-21	3.3
		Interaction	18372.9	4	4593.2	163.2	1.5e-22	2.6
		Within	1013.2	36	28.1	-	-	-
	/aba-/	Rate	3407.2	2	1703.6	69.3	4.5e-13	3.3
		Speakers	1803.9	2	902.0	36.7	2.1e-09	3.3
		Interaction	1334.0	4	333.5	13.6	7.7e-07	2.6
		Within	884.8	36	24.6	-	-	-
	/-acu-/	Rate	21236.3	2	10618.2	277.2	1.4e-22	3.3
		Speakers	25989.9	2	12995.0	339.3	4.4e-24	3.3
		Interaction	9060.1	4	2265.0	59.1	2.5e-15	2.6
		Within	1378.8	36	38.3	-	-	-
S14	/-icke-/	Rate	27389.0	2	13694.5	302.5	3.1e-23	3.3
		Speakers	11809.2	2	5904.6	130.4	3.2e-17	3.3
		Interaction	12822.2	4	3205.6	70.8	1.5e-16	2.6
		Within	1630.0	36	45.3	-	-	-
	/up/	Rate	62118.7	2	31059.4	753.5	4.2e-30	3.3
S14	/up/	Speakers	2988.0	2	1494.0	36.2	2.4e-09	3.3
		Interaction	930.5	4	232.6	5.6	1.2e-03	2.6
		Within	1484.0	36	41.2	--	-	-
	/-acka-/	Rate	51005.9	2	25503.0	504.7	4.6e-27	3.3
		Speakers	1796.3	2	898.2	17.8	4.3e-06	3.3
		Interaction	883.6	4	220.9	4.4	5.5e-03	2.6
		Within	1819.2	36	50.5	-	-	-

A.3 Two-Way ANOVA of Closure and Burst Durations in VCV Utterances

Table A.5 Means and square root of variances of stop closure and burst durations in VCV utterances with different vowels (FS = female speaker, MS = male speaker, var. = variance)

Duration parameter	Vowel type	Speakers										
		FS6	FS13	FS18	FS20	FS22	MS2	MS4	MS8	MS10	MS15	Over-all
Closure Mean (ms)	/a/	92.3	62.0	74.0	110.2	79.5	92.8	63.0	123.5	84.2	86.3	86.8
	/i/	84.3	77.0	89.2	94.5	93.8	75.3	74.0	131.3	83.5	130.0	93.3
	/u/	87.7	91.0	78.0	106.3	90.3	93.8	62.0	146.0	77.0	154.3	98.7
	Overall	88.1	76.7	80.4	103.7	87.9	87.3	66.3	133.6	81.6	123.6	92.9
Closure $\sqrt{\text{var}}$ (ms)	/a/	15.9	14.9	17.4	19.4	15.0	24.3	10.4	11.9	7.4	14.9	23.5
	/i/	16.0	12.2	15.5	20.9	26.6	19.4	13.2	23.2	21.1	42.2	28.9
	/u/	15.1	18.4	13.4	24.9	20.2	7.8	15.6	31.6	8.5	25.5	33.6
	Overall	15.1	18.9	16.0	21.6	20.8	19.5	13.6	24.2	13.4	40.2	29.2

Duration parameter	Vowel type	Speakers										
		FS6	FS13	FS18	FS20	FS22	MS2	MS4	MS8	MS10	MS15	Over-all
Burst Mean (ms)	/a/	46.0	44.7	42.8	50.3	27.2	45.0	31.8	37.3	43.0	30.8	39.9
	/i/	53.2	48.5	46.7	56.3	36.5	60.2	51.8	31.2	41.2	55.8	48.1
	/u/	50.7	46.3	50.2	54.7	27.2	57.8	52.2	17.5	46.2	45.3	44.8
	Overall	49.9	46.5	46.6	53.8	30.3	54.3	45.3	28.7	43.4	44.0	44.3
Burst $\sqrt{\text{var}}$ (ms)	/a/	41.2	23.9	34.0	41.9	11.9	34.8	18.9	24.8	33.9	19.9	28.7
	/i/	40.4	34.1	30.2	32.6	19.7	43.6	34.8	18.6	30.4	52.6	33.5
	/u/	35.9	29.9	28.2	31.4	12.4	37.4	28.8	46.9	33.0	23.6	31.8
	Overall	37.0	27.9	29.2	33.6	14.9	37.1	28.3	31.7	30.6	34.7	31.4

Table A.6 Means and square root of variances of stop closure and burst durations for different stops (5 female and 5 male speakers, FS = female speaker, MS = male speaker, var = variance)

Duration parameter	Stop type	Speakers										
		FS6	FS13	FS18	fS20	FS22	MS2	MS4	MS8	MS10	MS15	Over-all
Closure mean (ms)	/b/	95.7	90.3	81.0	130.7	81.3	95.3	66.7	151.7	78.0	119.0	99.0
	/d/	101.7	77.7	84.0	119.3	74.0	102.0	75.3	145.7	93.3	123.0	99.6
	/g/	97.3	74.0	79.7	98.7	67.0	79.0	56.0	114.7	88.0	116.0	87.0
	/k/	74.7	76.0	75.7	93.7	92.0	81.7	62.3	116.0	79.0	124.3	87.5
	/p/	89.0	84.0	94.3	100.3	96.0	102.0	71.7	141.0	82.3	145.7	100.6
	/t/	70.3	58.0	67.7	79.3	117.0	64.0	66.0	132.7	68.7	113.3	83.7
	Overall	88.1	76.7	80.4	103.7	87.9	87.3	66.3	133.6	81.6	123.6	92.9
Closure $\sqrt{\text{var}}$ (ms)	/b/	14.3	16.9	7.2	4.9	5.9	22.4	13.0	47.2	15.4	41.1	32.2
	/d/	7.6	24.8	8.0	24.7	2.6	2.0	11.9	5.9	15.6	53.5	29.1
	/g/	5.1	26.9	11.7	20.5	15.1	13.1	22.1	21.7	12.5	46.9	26.6
	/k/	7.6	16.8	13.6	16.2	19.9	8.6	15.3	7.9	10.4	26.5	22.7
	/p/	16.5	10.0	5.0	3.0	9.5	16.0	13.6	14.8	11.9	44.1	27.8
	/t/	9.1	11.5	33.6	9.7	22.7	22.5	3.6	12.6	9.1	58.0	32.9
	Overall	15.1	18.9	16.0	21.6	20.8	19.5	13.6	24.2	13.4	40.2	29.2
Burst mean (ms)	/b/	11.0	18.3	12.0	9.0	16.3	19.7	14.3	-17.3	11.0	14.7	10.9
	/d/	15.7	23.3	20.0	33.3	12.7	17.0	15.0	12.7	11.3	23.0	18.4
	/g/	18.3	23.3	30.3	37.0	34.0	36.7	38.3	31.0	23.3	25.0	29.7
	/k/	91.0	77.7	69.7	85.7	40.7	88.3	70.0	58.3	77.3	74.7	73.3
	/p/	74.0	56.0	63.0	67.3	37.0	53.3	62.0	42.7	59.7	38.3	55.3
	/t/	89.7	80.3	84.3	90.3	41.0	111.0	72.0	44.7	78.0	88.3	78.0
	Overall	49.9	46.5	46.6	53.8	30.3	54.3	45.3	28.7	43.4	44.0	44.3
Burst $\sqrt{\text{var}}$ (ms)	/b/	3.6	5.8	2.6	4.4	3.2	14.2	6.5	47.3	1.7	8.1	16.9
	/d/	7.0	9.3	3.6	30.2	3.8	3.6	1.0	3.5	2.9	4.4	10.9
	/g/	8.1	3.5	12.6	20.7	24.3	13.9	10.8	11.5	12.7	6.0	13.2
	/k/	8.0	16.4	15.3	8.7	2.5	12.4	16.5	4.7	9.6	18.8	17.8
	/p/	2.6	6.6	13.0	16.6	7.2	4.9	30.5	22.2	5.7	10.6	17.1
	/t/	11.1	6.5	8.7	4.0	5.3	14.1	12.5	8.1	1.7	52.5	25.6
	Overall	37.0	27.9	29.2	33.6	14.9	37.1	28.3	31.7	30.6	34.7	31.4

REFERENCES

[1] S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460-475, April 1976.

[2] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept. 1997.

[3] D. O'Shaughnessy, *Speech Communication – Human and Machine*, 2nd ed. Hyderabad: Universities Press, 2001, pp. 56-107, 173-214, 437-459.

[4] G. Doddington, "Speaker recognition - identifying people by their voices," *Proc. IEEE*, vol. 73, no. 11, pp. 1651-1664, Nov. 1985.

[5] M. Fraundez-Zanuy and E. Monte-Moreno, "State-of-the-art in speaker recognition," *IEEE A&E Systems Magazine,* pp. 7-12, May 2005.

[6] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Comm.*, vol. 50, pp.782-796, June 2008.

[7] G. R. Gangula / Prof. P. C. Pandey (supervisor), "Study of speech analysis parameters for speaker recognition," *M.Tech. dissertation*, Department of Electrical Engineering, Indian Institute of Technology, Bombay, July 2006.

[8] S. Furui, "Recent advances in speaker recognition," in *Proc. Int. Conf. Audio- and Video-based Biometric Person Authentication* 1997, vol. 1206, pp. 237–252.

[9] R. Wildermoth, "Text-independent speaker recognition using source based features," M. Phil. thesis, Griffith University, Australia, Jan. 2001.

[10] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.

[11] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578-489, Oct. 1994.

[12] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture models," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, Jan. 1995.

[13] R. Zheng, S. Zhang, B. Xu, "A comparative study of feature and score normalization for speaker verification," in *Advances in Biometrics*, Heidelberg: Springer Berlin, 2005, pp. 531-538.

[14] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Comm.*, vol. 25, pp. 133-147, 1998.

[15] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[16] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 2, pp. 254-272, Apr. 1981.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[18] K. Yu, J. Mason, and J. Oglesby, "Speaker recognition using hidden Markov models, dynamic time warping and vector quantization," *IEE Proc. Vision, Image and Signal Processing*, vol. 142, no. 5, pp. 313-318, Oct. 1995.

[19] G. R. Gangula, P. C. Pandey and P. K. Lehana, "Application of harmonic plus noise model for enhancing speaker recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 3040-3041, Nov. 2006.

[20] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-285, Feb. 1989.

[21] M. R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-23, no. 2, pp. 176-182,Apr. 1975.

[22] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, April 1979.

[23] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey Conf. 2001*, pp. 213-218.

[24] J. Makhoul, "Linear Prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561-579, April 1975.

[25] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Delhi: Pearson Education, 1993, pp. 371-388, 412-432.

[26] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, Apr. 1990.

[27] B. Milner, "A comparison of front-end configurations for robust speech recognition," in *Proc. ICASSP* 2002, vol. 1, pp. 797- 800.

[28] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 639-643, Oct. 1994.

[29] Chow and W. H. Abdulla, "Robust speaker identification based on perceptual log area ration and Gaussian mixture models," in *Proc. Interspeech-2004*, pp. 1761-1764.

[30] J. W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, no. 9, pp. 1215-1247, Sept. 1993.

[31] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 21-29, Jan. 2001.

[32] M. A. Kohler, W. D. Andrews, J. P. Campbell, and J. H. Hernandez-Cordero," Phonetic speaker recognition," in *Proc. IEEE Signals, Systems and Computers* 2001, vol. 2, pp. 1557-1561.

[33] D. A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. ICASSP* 2003, vol. 4, pp. 784–787.

[34] E. Shriberg, L.Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic sequences for speaker recognition," in *Speech Communication*, vol. 46, pp. 455-472,July 2005.

[35] L. Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, "Modeling duration patterns for speaker recognition," in *Proc. Eurospeech* 2003, pp. 2017-2020.

[36] A. M. Dean and D. T. Voss, *Design and Analysis of Experiments*, New York: Springer-Verlag, 1999, pp. 33-49, 135-168.

[37] *NIST/SEMATECH e-Handbook of Statistical Methods*, July 18, 2006. [Online]. Available: http://www.itl.nist.gov/div898/handbook/prc/section4/prc4.htm. [Accessed: July 23, 2009].

[38] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *J. Acoust. Soc. Am.,* vol. 100 (5), pp. 3417-3430, Nov. 1996.

[39] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," *J. Acoust. Soc. Am*., vol. 115 (3), pp. 1296-1305, Mar. 2004.

[40] A. M. A. Ali, J. V. Spiegel, and P. Mueller, "Acoustic–phonetic features for the automatic classification of stop consonants," *IEEE Trans. Speech and Audio Processing*, vol.9, no. 8, pp. 833-841, Nov. 2001.

[41] A. W. Howitt, "Automatic syllable detection for vowel landmarks" Sc. D. thesis, Massachusetts Institute of Technology, USA, 2000.

[42] A. R. Jayan, P. C. Pandey, and P. K. Lehana, "Time-scaling of consonant-vowel transitions using harmonic-plus-noise model for improving speech perception by listeners with moderate sensorineural impairment" in *Proc. 19th Int. Congress on Acoustics* 2007, paper no. CAS-03-006.

[43] P. Mermelstein, "Acoustic segmentation of speech into syllabic units", *J. Acoust. Soc. Am.*, vol. 58, no. 4, pp. 880-883, Feb. 2002.

[44] P. Niyogi and M. Sondhi, "Detecting stop consonants in continuous speech," *J. Acoust. Soc. Am.*, vol. 111, no. 2, pp. 1063-1076, Feb. 2002.

[45] P. Niyogi, C. Burges, and P. Ramesh, "Distinctive feature detection using support vector machines,**"** in *Proc.ICASSP* 1999, vol. 1, pp. 425-428.

[46] A. R. Jayan and P. C. Pandey, "Detection of stop landmarks using Gaussian mixture modeling of speech spectrum," in *Proc. ICASSP* 2009, pp. 481-484.

[47] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech and signal Processing*, vol. ASSP-28, no. 4, pp.357-366, Aug. 1980.

[48] T. D. Bui and G. Chen, "Translation-invariant denosing using multiwavelets," *IEEE Trans. Signal Processing*, vol. 46, no. 12, Dec. 1998.

[49] R. R. Coifman and D. L. Donoho, "Translation invariant de-noising," in *Wavelets and Statistics, Springer Lecture Notes in Statistics*, New York: Springer-Verlag, 1994, vol. 103, pp. 125–150.