Speaker Identification and Verification Using Line Spectral Frequencies

Pujita Raman

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

In

Electrical Engineering

Dr. A. A. (Louis) Beex, Chair

Dr. William T. Baumann

Dr. Guoqiang Yu

May 4, 2015

Blacksburg, Virginia

Keywords: Speech, Speaker, Noise, Identification, Verification, Recognition, Feature, Line Spectral Frequency, Gaussian Mixture Model, Transition, Vowel

Speaker Identification and Verification Using Line Spectral Frequencies

Pujita Raman

ABSTRACT

State-of-the-art speaker identification and verification (SIV) systems provide near perfect performance under clean conditions. However, their performance deteriorates in the presence of background noise. Many feature compensation, model compensation and signal enhancement techniques have been proposed to improve the noise-robustness of SIV systems. Most of these techniques require extensive training, are computationally expensive or make assumptions about the noise characteristics. There has not been much focus on analyzing the relative importance, or 'speaker-discriminative power' of different speech zones, particularly under noisy conditions.

In this work, an automatic, text-independent speaker identification (SI) system and speaker verification (SV) system is proposed using Line Spectral Frequency (LSF) features. The performance of the proposed SI and SV systems are evaluated under various types of background noise. A score-level fusion based technique is implemented to extract complementary information from static and dynamic LSF features. The proposed score-level fusion based SI and SV systems are found to be more robust under noisy conditions.

In addition, we investigate the speaker-discriminative power of different speech zones such as vowels, non-vowels and transitions. Rapidly varying regions of speech such as consonant-vowel transitions are found to be most speaker-discriminative in high SNR conditions. Steady, high-energy vowel regions are robust against noise and are hence most speaker-discriminative in low SNR conditions. We show that selectively utilizing features from a combination of transition and steady vowel zones further improves the performance of the score-level fusion based SI and SV systems under noisy conditions.

ACKNOWLEDGEMENTS

I would like to thank my graduate advisor, Dr. A. A. (Louis) Beex, for giving me an opportunity to work on such an interesting research problem at the DSPRL. His constant encouragement and guidance over the past two years has been the driving force behind the successful completion of this thesis. With his wonderful insights and ideas, he has taught me to never give up, and approach problems from a new perspective, even when there seems to be no silver lining in sight. I also thank Dr. William T. Baumann and Dr. Guoqiang Yu for being a part of my thesis committee.

None of this would have been possible without the support of my parents, grandparents and very large family. Thank you all for your unconditional love, incessant jokes and inspiring words of wisdom. Special thanks to my sister, Atmaja and my dog, Sirius, for being experts at cheering me up, even in the worst of times. I am deeply indebted to my best friend and my soulmate, Girish, for standing by me every step of the way, and showing me that love knows no distance. I am grateful to all my amazing friends, for making Blacksburg feel like home away from home.

I dedicate this thesis to my grandfather, Itha. Words cannot express how much I miss him. He was, he is, and he will always be my hero.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AR	Auto-regressive
CV	Consonant Vowel
DET	Detection Error Trade-off
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
EER	Equal Error Rate
EM	Expectation Maximization
FIR	Finite Impulse Response
FLF	Feature-level Fusion
FOGD	First Order Gaussian Differentiator
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Model
HE	Hilbert Envelope
IA	Identification Accuracy
IR	Identification Rate
LLR	Log-likelihood Ratio
LP	Linear Prediction
LSF	Line Spectral Frequency
MAP	Maximum a posteriori
MFCC	Mel Frequency Cepstral Coefficients
MR	Miss Rate
RMSD	Root Mean Square Deviation
ROC	Receiver Operator Characteristic
SI	Speaker Identification
SIV	Speaker Identification and Verification
SLF	Score-level Fusion
SNR	Signal to Noise Ratio
SPE	Spectral Peaks Energy
SPIB	Signal Processing Information Base
SR	Spurious Rate
SV	Speaker Verification
TIMIT	Texas Instruments – Massachusetts Institute of Technology
UBM	Universal Background Model
VAD	Voice Activity Detection
VEP	Vowel End Point
VOP	Vowel Onset Point
ZFFS	Zero Frequency Filtered Signal

1 INTRODUCTION

Speech is the primary form of human communication. From at least a 100,000 years ago, humans have been able to express their thoughts and emotions by stringing together different permutations of vowels and consonants. While the same words might be spoken by different speakers, no two individuals sound identical because of differences in the shape and size of their vocal apparatus. In fact, speech contains underlying information about the identity, gender, health, ethnicity, and even the emotional state of the speaker.

1.1 SPEAKER RECOGNITION

Speaker recognition refers to the task of determining a speaker's identity using information extracted from his/her voice. It is also referred to as voice recognition or voice biometrics in some literature. When human listeners infer a speaker's identity, it is known as *auditory speaker recognition* [1]. In *semi-automatic speaker recognition*, human experts decide the speaker's identity with the aid of machines, by analyzing various acoustic parameters such as resonant frequencies, spectral energy etc. When the speaker recognition task is performed entirely by a machine, without any human aid, it is termed *automatic speaker recognition*. In this case, the machine must learn the characteristics of each speaker, by extracting speaker-specific information known as *'features'* from useful segments of the speech signal, and suppressing statistically redundant information from the non-useful segments [2].

1.2 SPEECH PRODUCTION IN HUMANS

In order to understand what speaker-specific information is present in a speech signal, we must first analyze how speech is produced in humans. In terms of functionality, the human vocal apparatus can be broadly divided into four sections: an air source (lungs), a vibrating component (vocal folds/cords in the larynx), resonant chambers (pharynx, mouth and nasal cavities) and articulators (tongue, palate, cheek, teeth, and lips) [3]. The human vocal system is depicted in Figure 1.

The air required for speech production originates in the lungs, creating an air stream that passes through the trachea to the larynx. The larynx, also known as the 'voice box', is located on top of the trachea and contains two folds known as *vocal cords*. These opening between the vocal cords is known as the *glottis*. The air stream that originates in the lungs passes between the vocal cords, causing a pressure drop across the larynx.

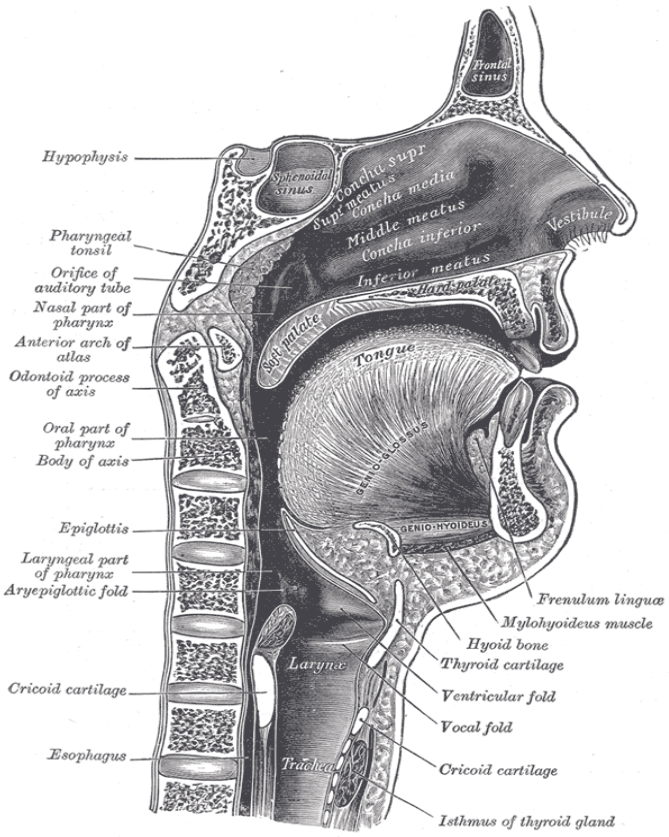

Fig. 1.1: The human vocal apparatus.[1]

If the pressure drop is sufficient to separate the vocal cords, they draw apart, and then close immediately due to their tension, elasticity, and the Bernoulli effect [4]. This causes the air pressure beneath the glottis to increase again, and the process repeats. Due to the rapid opening and closing of the glottis, the vocal cords release the air from the lungs in a vibrating stream. This is known as *phonation*, and the resulting speech is called *voiced* speech. The muscles of the larynx adjust the length and tension of the vocal cords, which in turn affects the fundamental frequency of vibration of the vocal cords (directly related to *pitch*) [5]. If the pressure across the larynx is not sufficient, or if the vocal cords are not under sufficient tension, the vocal cords will not oscillate. This is known as *unvoiced* speech.

All the resonant chambers above the vocal cords constitute the human *vocal tract*. This includes the laryngeal pharynx, oral pharynx, oral cavity, nasal pharynx, and nasal cavity [4]. When the signal generated passes through the vocal tract, its frequency content is altered by the resonance properties of the vocal tract. These resonant frequencies of the vocal tract are called *formants*. Thus, the shape of the vocal tract can be estimated from the spectral shape of the speech signal. In

[1] Image source: http://upload.wikimedia.org/wikipedia/commons/2/20/

addition, the vocal tract can close, constrict and change its shape in different ways, resulting in different kinds of sounds.

The articulators, consisting of the tongue, palate, cheek, teeth, and lips, further modulate the sound emanating from the larynx, by obstructing or allowing airflow through the mouth in a number of ways [6]. Thus, the vocal cords, vocal tract, and articulators work in harmony during speech production, leading to an enormous array of different sounds.

The elementary sounds that occur in a language are called *phones*. Different phones might correspond to the same linguistic unit, called a *phoneme*. For example, the 'p' sound in the word 'pit' involves the drawing of breath (aspiration) and is denoted by [p^h]. In contrast, the 'p' sound in 'spill' is unaspirated and is denoted by [p]. However, both these phones correspond to the same phoneme /p/. When a phone is pronounced with an open vocal tract, without occluding the air coming from the glottis, it is called a *vowel*. In contrast, when a phone is produced by a completely or partially closed vocal tract, it is called a *consonant*.

1.3 SPEAKER INDIVIDUALITY

Although the basic speech production mechanism is the same in all humans, it is influenced by many physiological properties of the speaker. Some of these properties include vocal tract shape, lung capacity, maximum phonation time, and glottal air flow [4]. While these physiological properties are not directly measurable from speech, their effects are felt in other quantifiable acoustic properties. For example, the pitch contains information about the glottis, and the formants contain information about the vocal tract and nasal cavities.

Thus, the underlying speaker-specific information in speech can be broadly classified into low-level features and high level features. *Low-level features* are related to the physiological and acoustic aspects of a person's vocal apparatus, such as spectral energy, formant frequencies and bandwidth. These features can be easily measured from a speech signal. On the other hand, *high-level features*, also called behavioral or learned traits, refer to the semantic and linguistic aspects of speech such as prosody, speaking style, vocabulary, accent, and pronunciation [7]. These features are difficult to extract from a speech signal and quantify. While the human brain effortlessly uses a combination of these subtle high-level as well as low-level features to determine a speaker's identity, this task is not that easy for a machine to accomplish. Since low-level features are easy to measure, they are more widely used in automatic speech recognition.

1.4 AUTOMATIC SPEAKER RECOGNITION SYSTEMS

An automatic speaker recognition system is built and tested in two phases: *enrollment/training* phase, and *recognition* phase. In order to build a speaker recognition system, a speaker model or voice-print must be created for each of the candidate speakers, and stored in the system database. This is called the training or enrollment phase, and is depicted in Fig. 1.2. In this phase, a feature

extraction module estimates a set of feature vectors which encapsulate speaker-specific information, from the speech signal. These feature vectors are then used to build a statistical model pertaining to that speaker. In this way, models are trained and stored for each of the candidate speakers during the enrollment phase.

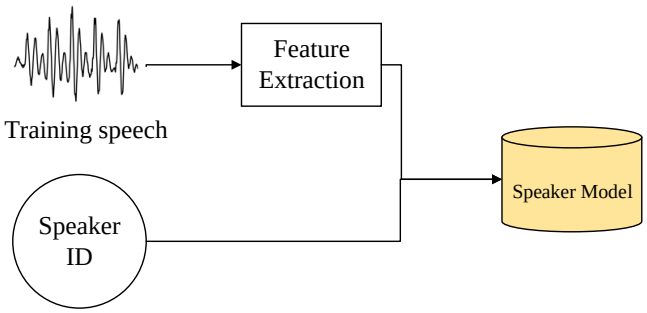

Fig. 1.2: Evaluation phase of a speaker recognition system.

In the recognition phase, feature vectors are obtained from the unknown speaker's test utterance using the same feature extraction process. Then, the feature vectors are compared against the model(s) in the system database. The pattern matching algorithm assigns a score for each feature vector – this score tells us how similar the feature is with the model under consideration. The similarity scores for all the feature vectors from the test utterance are combined to make a decision.

Furthermore, the text used for training and testing a speaker recognition system can be constrained to a specific word/phrase (*text-dependent*), or a set of words/phrases (*text-prompted*, *or vocabulary-dependent*) or be completely unconstrained (*text-independent*) [7]. A text-dependent speaker recognition system is very vulnerable, as it can be fooled by recording the claimant and playing it to the system.

Speaker recognition encompasses two fundamental tasks, namely, *speaker identification* and *speaker verification*. In speaker verification, the unknown speaker first claims his identity using an ID card or a username/password, and then requests voice-based authentication. Given an utterance of speech, speaker verification is the task of determining whether the unknown speaker is who he/she claims to be. Thus, verification is a 2-class decision task, where the two classes are 'claimed speaker' and 'impostor speaker' [8].

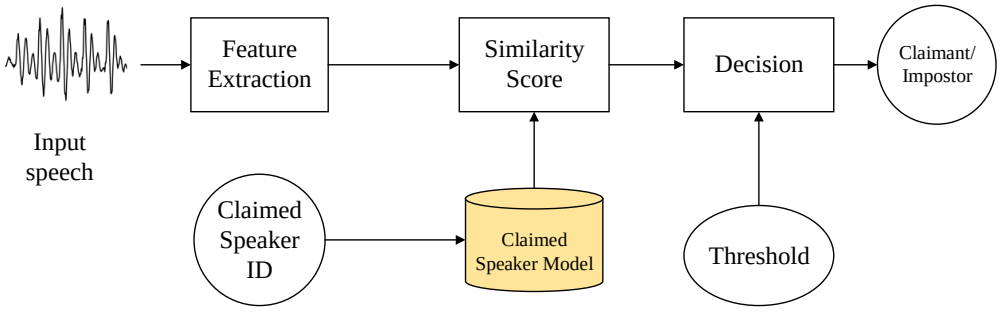

Fig. 1.3: A speaker verification system.

The block diagram of a typical speaker verification system is shown in Fig. 1.3. During verification, feature vectors are estimated from the input speech and compared to the claimed speaker model to determine a similarity score. If this score is above a certain threshold, the unknown speaker's identity claim is declared to be true. Else, he/she is declared to be an impostor.

In contrast, in speaker identification, the unknown speaker makes no such *a priori* identity claim. Given an utterance of speech, speaker identification is the task of determining who among a set of candidate speakers said it. When the actual speaker is always from the set of candidate speakers, it is termed *closed-set speaker identification*. Closed-set speaker identification is an N-class decision task, where N is the number of candidate speakers [8]. However, in *open-set speaker identification*, the actual speaker might not be from the set of candidate speakers. In this case, an additional decision, "the unknown speaker does not match any of N speakers" is required.

The block diagram of a closed-set speaker identification system is shown in Fig. 1.4. During speaker identification, feature vectors are extracted from the input speech and compared with each speaker model to determine a similarity score. The unknown speaker's identity is decided to be the speaker whose similarity score is the highest among all the candidate speakers.

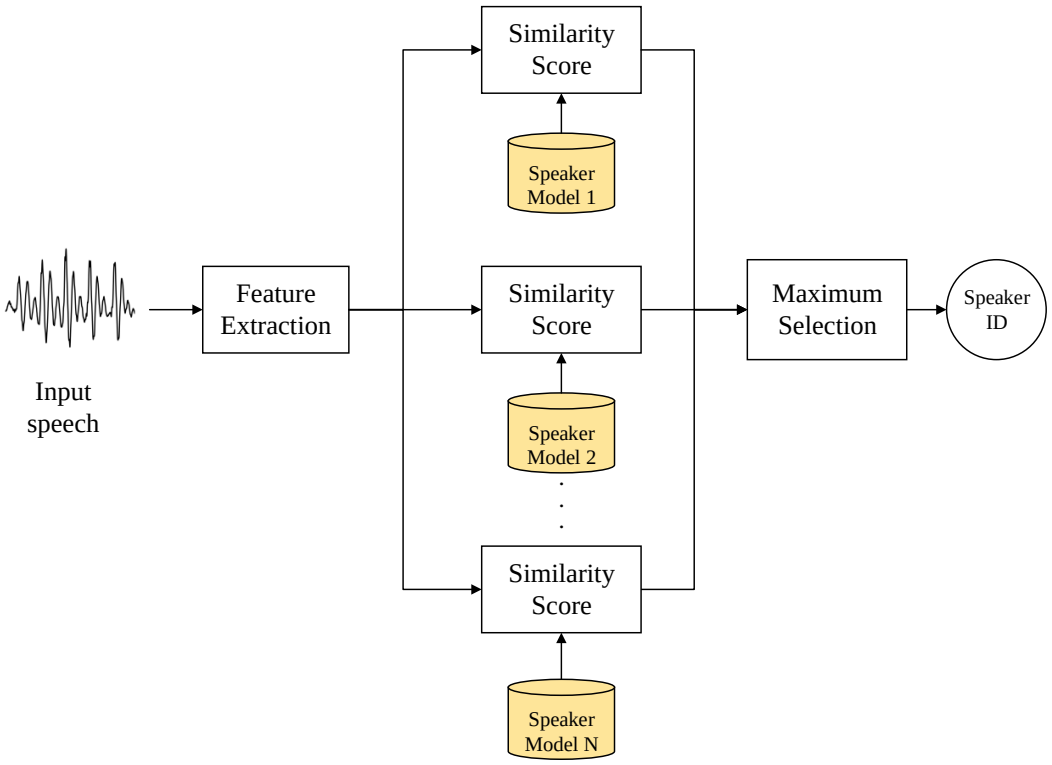

Fig. 1.4: A speaker identification system.

1.5 APPLICATIONS OF SPEAKER RECOGNITION

Speaker recognition has widespread application in a number of areas such as:

1. *Physical Access Control:* While other biometric technologies require special equipment such as fingerprint or retinal scanners, only a microphone is needed to capture a speech signal. In addition, since speech is a natural form of communication, it feels non-intrusive from a user's point of view. Thus, speaker recognition is a low-cost, non-contact biometric technology that can be used for authentication and physical access control to secure zones such as in airports and offices. Speaker verification can be incorporated as an additional security layer in existing access control systems to confirm a person's identity.

2. *Remote Access Control:* When compared to other biometric technologies, speaker recognition is more suitable for remote access control scenarios where the user is not physically present at the location. For example, speaker recognition can be used to authenticate remote access of databases or computer networks.

3. *Telephone Security:* Under low noise conditions, automatic speaker recognition systems are capable of very good performance. Thus, speaker recognition is being used by banks such as Barclays Wealth to provide fast voice-based authentication in telephone banking services and credit card transactions [9].

4. *Metadata Extraction:* As the volume of information has been increasing exponentially over the years, it is very crucial to extract metadata for quick searching and indexing of audio streams. Speaker recognition can be used for automatic detection and labeling of speakers in audio/video documents such as teleconference meetings, videos, and TV broadcasts [2].

5. *Surveillance:* Speaker recognition technology can also be used to detect and track persons of interest during telephone/radio surveillance by security agencies [10].

6. *Forensics:* Forensics is another important application of speaker recognition. For example, a suspect's involvement in a crime can be verified by running speech samples that were recorded during the commitment of the crime through a recognition system.

7. *Personalized Technology:* The signal processing and pattern recognition algorithms for speaker recognition are low-cost and memory-efficient, and thus can be easily ported to even mobile devices [11]. Thus, speaker recognition can be used in personalized user interfaces, smart cars, games, and smartphone assistants such as Siri.

1.6 TECHNICAL CHALLENGES IN SPEAKER RECOGNITION

There are a number of practical difficulties that must be overcome when building a speaker recognition system, some of which are outlined below:

1. *Intra-speaker variability:* In real-world speaker recognition systems, training and testing does not usually happen in the same session. A speaker's voice changes over time, as a result of which there might be a lot of inter-session variability [8]. The speaker's health condition and emotional state also leads to variations in his/her voice. Other factors include changes in speech effort levels and variations in speaking rate [8]. Thus, a real-world speaker recognition system must be robust to intra-speaker variability.

2. *Large Population Sizes:* In a finite feature space, as the number of enrolled speakers increases, the recognition performance decreases due to speaker distribution overlap [12]. Thus, the feature extraction as well as the speaker modeling algorithms selected must be able to *maximize inter-speaker variability*, enabling good performance even over very large speaker sets.

3. *Immunity to Spoofing Attacks:* The speaker-recognition system must not be vulnerable to attacks by impostors who try to mimic the voice of a candidate speaker.

4. *Channel Distortion:* In speaker recognition systems, the speech might be distorted, and interference might be introduced during its transmission through the communication channel. In addition, the microphone/channel type might also vary from one session to another, for example, during telephone based recognition. This is known as a mismatched condition, and is one of the most serious sources of error in speaker recognition [11].

5. *Noisy Environments:* In real life scenarios, the speech signal will be affected by varying degrees of noise (from traffic, background speakers, echoing etc.), which will mostly be non-stationary. Thus, the speaker recognition system must be robust to noisy environments.

6. *Short Test Utterances:* Since users would not be willing to speak for long durations of time, a speaker recognition system must be able to provide good performance even with short test utterances.

7. *Quick Learning:* The training procedure should be user-friendly and not be too time-consuming. The ad-hoc enrollment of a new speaker into the system should also be easy to accomplish.

8. *Fast Response:* The speaker recognition system must be able to provide high performance with very minimal delay, i.e. it must operate in what feels like real-time.

1.7 MOTIVATION AND OUTLINE

As the applications of speaker recognition keep growing, there is an increasing need to improve the performance of SIV systems that operate in noisy and/or otherwise distorting conditions commonly encountered in the real world. The aim of this thesis is to develop automatic, text-independent speaker identification (SI) and verification (SV) systems using Line Spectral Frequency features. During the course of this research, we analyze the performance of the SIV systems in the presence of different types of noise, and explore the fusion of static and dynamic LSF information to improve the robustness of the systems.

As mentioned in the previous section, a crucial ingredient for a SIV system is the extraction of features that can effectively characterize each speaker. Consequently, a question arises as to which parts of a speech signal contain important speaker-specific information that should be extracted, and which parts are not so relevant. In this thesis, we attempt to study the relative speaker-discriminative power of three zones of the speech signal, namely transitions, vowels and non-vowels, for use during the speaker recognition process. We hypothesize that the rapidly varying transition regions in speech are relatively more important compared to the steady vowel regions. In particular, we are interested in transitions into and out of vowels. We investigate the effect of noise on the speaker-discriminative power of these transition speech zones. Subsequently, we analyze whether selectively utilizing features from only the speaker-discriminative zones, and discarding less relevant features during the testing/recognition phase benefits the performance of the SIV systems.

The rest of the thesis is organized as follows. Chapter 2 contains a review of the previous research in automatic text-independent speaker recognition, focusing on feature extraction, speaker modeling, noise-robust speaker recognition and the importance of speech zones. In Chapter 3, an overview of the feature extraction process used in this thesis is provided. A review of the techniques used for speaker modeling in our speaker identification and speaker verification systems is given in Chapter 4. Chapter 5 contains a description of the method used to localize the transitions into and out of vowels. In Chapter 6, we discuss the various tests and simulations performed in speaker identification, along with the results. In Chapter 7, we describe the various tests and simulations performed in speaker verification, along with the results. In Chapter 8, we draw conclusions based on our results and explore future research possibilities.

2 BACKGROUND AND MOTIVATION

In this chapter, we review the major advances in speaker recognition technology over the past 50 years, focusing on feature extraction and speaker modeling techniques. In parallel, we also present the motivations for our research.

2.1 FEATURE EXTRACTION

The earliest known research on speaker recognition dates back to the 1950s. These papers investigated the effect of different speech characteristics such as speaking rate, duration, and frequency content, on speaker recognition by human listeners [13, 14]. The first attempt to build an automatic speaker recognition system was made by Pruzansky at Bell Labs in 1963 [15]. The speech was converted into time-frequency energy patterns, known as spectrograms, and recognition was performed by cross-correlating these patterns with reference patterns and measuring their similarity. Texas Instruments built a prototype system that was tested by the U.S. Air Force and The MITRE Corporation in 1976 [9]. Around this time, Bell Labs also built experimental speaker recognition systems aimed to work over telephone lines [16]. From the 1970s until today, a number of different features have been investigated for use in speaker recognition systems [4, 16].

Various features can be extracted from a speech signal, but only those which exhibit high inter-speaker variability and low intra-speaker variability are useful in speaker identification and verification (SIV). Features used in SIV should be easy to measure and difficult to impersonate. They should also be robust against noise, distortion, and variations in the voice or health of the speaker [2]. The size of the features also affects the performance of the speaker recognition system. The features that are used in speaker recognition can be broadly classified into low-level features and high level features.

2.1.1 Low-level Features

Low-level features are related to the physiological and acoustic aspects of a person's vocal apparatus, such as spectral energy, formant frequencies, and bandwidth. These features are typically extracted from short segments of speech. Although low-level features can be easily measured, they are known to be sensitive to noise and channel mismatch conditions [17].

As described in Section 1.3, when we speak, each phone is produced by a different configuration of our vocal apparatus, and thus has a different frequency content. Since the frequency content of a speech signal changes as we move from one phone to another, it is considered to be *non-stationary* in nature. The spectrogram of a speech signal is shown in Fig. 2.1. From the figure, we can clearly see that the frequency content of the signal varies over time. In addition, the energy of

the speech signal is concentrated around certain frequencies. These frequencies are known as formants, and correspond to the resonant frequencies of the vocal tract.

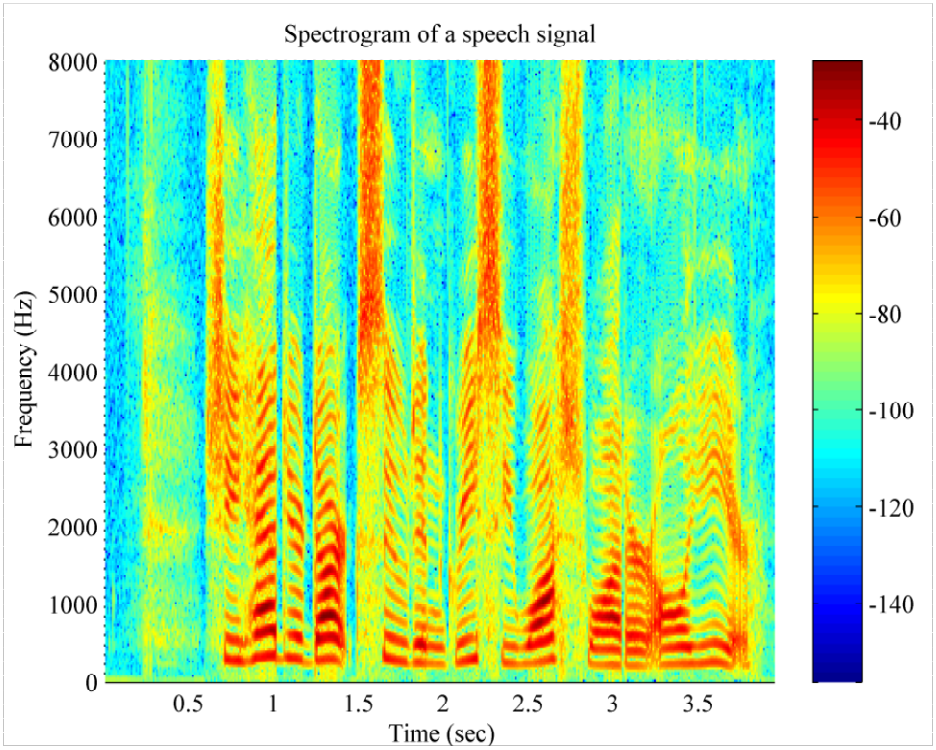

Fig. 2.1: Spectrogram of a speech signal.

Most signal processing techniques are based on the assumption that the signal under consideration is stationary. Hence, before low-level feature extraction, the speech signal is split into short, overlapping segments called frames, within which it is assumed to have *quasi-stationary* behavior. A frame is usually 10-30 milliseconds long, with around 30-50% of overlap between frames commonly used. Overlapping is performed to ensure temporally smoother parameter transitions between frames. Before feature extraction, the frame is also usually passed through a first-order, high pass filter to boost higher frequencies and counter the downward sloping of the speech spectrum. This process is known as pre-emphasis. The frame is also multiplied by a smooth window function to taper the frame ends and avoid discontinuity at the edges.

Low-level features can be broadly classified into spectral features and voice source features. These categories are explained in the following sections.

2.1.1.1 Spectral Features

Spectral features characterize the spectral envelope of the speech signal, which contains information about the resonance properties of the vocal tract [11].

One of the most popular spectral features used in speech and speaker recognition is the Mel Frequency Cepstral Coefficient (MFCC). Introduced in 1980 by Davis and Mermelstein [18],

MFCCs are a representation of the short-term power spectrum of a speech signal. Using a triangular filter bank, as shown in Fig. 2.2, the speech spectrum is first mapped from a linear scale onto a mel-scale, in order to approximate the human auditory response more closely [19]. A mel-scale is a special scale based on the perceived pitch of a tone. This mapping is followed by logarithmic compression and discrete cosine transform (DCT) to obtain the MFCCs. A number of different MFCC implementations exist today, differing mainly in the number of filters, filter characteristics, and in the manner in which the spectrum is warped [20].

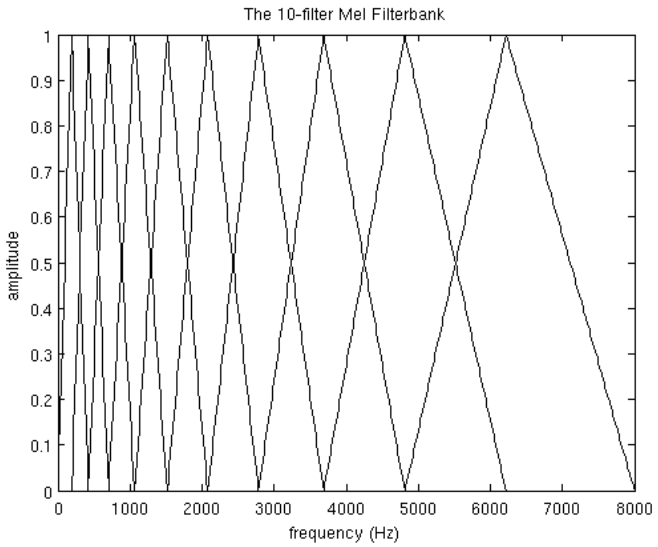

Fig. 2.2: A mel filter bank [2].

While MFCCs are the most commonly used features in speaker recognition, the presence of background noise or channel distortions lowers the performance of MFCC based methods significantly [21]. Recently, a number of alternative features have been proposed by modifying the cepstral feature extraction process to make it more noise resistant. Some of these features are Mean Hilbert Envelope Coefficient (MHEC), Power Normalized Cepstral Coefficient (PNCC), Normalized Modulation Cepstral Coefficient (NMCC), Minimum Variance Distortionless Response (MVDR), Spectral Centroid Frequency, and Magnitude (SCF, SCM), and Multitaper MFCC [19, 21]. Combining the MFCC information with phase information has also resulted in improved recognition performance in the presence of noise [22, 23].

Linear Prediction (LP), used commonly in speech coding, is another short-term spectral estimation technique that is also used in speaker recognition applications [24]. Linear prediction is based on the source-filter model of speech production. A quasi-stationary speech frame is considered to be the output of an autoregressive (AR) system (which represents the vocal tract) driven by an

[2] Image source: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

excitation. The excitation represents the glottal pulse when the speech is voiced, and white noise when the speech is unvoiced. The vocal tract is modeled as an all-pole filter, whose coefficients are called the Linear Prediction Coefficients (LPC). The poles of this filter correspond to the resonant frequencies of the vocal tract, i.e. the formants.

The predictor coefficients themselves are rarely used as features, but they are transformed into more robust features such as formant frequencies and bandwidths, Reflection Coefficients (RC), Log Area Ratios (LAR), Line Spectral Frequencies (LSF), Linear Predictive Cepstral Coefficients (LPCCs), and Perceptual Linear Prediction (PLP) coefficients. These equivalent parameterizations have different properties. For example, all LPC coefficients change when one of the formants changes, but in contrast, the reflection coefficients at lower model order are not affected by going to a higher model order. Frequency domain linear prediction (FDLP) [25] and Temporally Weighted Linear Prediction [26] have also been suggested to improve speaker recognition performance on noisy and reverberant speech.

Among these different LPC representations, the LSF decomposition, first introduced by Itakura [27], has gained the most popularity, with applications in speech coding, speech recognition [28], as well as speaker recognition [29]. The LP polynomial is decomposed into two polynomials, each representing one resonance condition of the vocal tract, i.e. fully open or fully closed at the glottis [30]. The roots of these polynomials are the set of Line Spectral Frequencies. LSFs have been shown to be more efficient than other representations for encoding LPC spectral information [29], as they can be quantized with fewer bits, while retaining the same speech quality. Also, they always result in a stable LPC filter, even under severe quantization noise. This stability feature even allows the LSFs to be interpolated across frames [30].

In clean conditions, very good recognition rates have resulted when LSFs and its various transformations were used as features in text-independent speaker recognition systems [29, 31]. Perceptual Line Spectral Frequencies (PLSF) [32] and Mel Line Spectral Frequencies (MLSF) [33] have also been investigated as features for speaker identification. The robust nature of the LSF representation hints at its potential as a feature for SIV in noisy conditions. However, this hypothesis does not seem to have been investigated until now. In our research, we have chosen the LSF representation for feature extraction, explained in detail in Chapter 3.

Given the non-stationary nature of speech, temporal variations such as formant transitions and energy modulations contain very useful speaker-specific information. In order to incorporate more dynamic/transitional information, 1st and 2nd order spectro-temporal features, known as *delta* and *delta-delta* features are sometimes utilized [34, 35]. They may be computed as the difference between adjacent spectral feature vectors [36], or as regression of adjacent feature vectors [35] , and are appended with the static feature vectors for each frame. The delta LSF representation has proved to be useful in voice conversion, speech recognition [28], and emotion recognition [37], but has not been used extensively in speaker recognition.

2.1.1.2 Voice Source Features

Voice source features characterize the glottal source of voiced sounds known as the glottal volume velocity waveform, or simply the glottal flow [38]. The derivative of the glottal flow exhibits a negative impulse-like response at the glottal closure instants (GCI). This is called the glottal pulse, and serves as the primary excitation signal during speech production.

One of the most popular voice source features is the rate of vibration of the vocal folds, known as the fundamental frequency (F_0) [2]. Other voice source features mainly describe the glottal pulse shape and the glottal closure duration. In order to estimate the glottal pulse, the glottal source and the vocal tract are assumed to function independent of each other. Using this assumption, we can first model the vocal tract as an all-pole filter using linear prediction. Then, we can estimate the excitation signal (glottal pulse) by filtering the speech signal using the inverse of the vocal tract model. The resulting signal is called the LP residual, and this process is called glottal inverse filtering. The periodic peaks in the LP residual occur at the glottal closure instants. The glottal flow derivative can be parameterized by fitting physical glottal flow models to the inverse filtered signal [39]. Other approaches include using wavelet smoothing excitation [40], Adaptive Forced Response Inverse Filtering (AFRIF) [41], Iterative adaptive inverse filtering (IAIF) [38], and closed-phase covariance analysis [42]. The LP residual, by itself, has also been used for feature extraction using AANN [43, 44]. Glottal Flow Cepstral Coefficients (GFCC) and Residual Phase Cepstral Coefficients (RPCC) have also been proposed [45].

2.1.2 High-level Features

In recent years, there has been a lot of focus on using high-level features in speaker recognition. High-level features, also called behavioral or learned traits, refer to the semantic and linguistic aspects of speech such as prosody, speaking style, vocabulary, accent, and pronunciation. High-level features usually arise from longer segments of speech such as words or phones. While high-level features are more difficult to measure when compared to low-level features, they are less affected by noise and channel mismatch. This is because people are less likely to change their idiosyncrasies due to noise or environmental change [17]. High-level features can supply complementary information to the low-level features and help improve overall recognition accuracy.

The idea behind high-level modeling is to convert each utterance into a sequence of tokens. These tokens could be words, phones, pitch gestures, etc. The tokens are a discrete representation of the input speech. The baseline classifier for token features is commonly an N-gram model, which is constructed by estimating the joint probability of N consecutive tokens. Recently SVM has also been proposed as an alternate language modeling approach for high-level speaker recognition [46]. High-level features can be broadly classified into prosodic, phonetic, and lexical features.

2.1.2.1 Prosodic Features

Prosodic features, i.e. features based on the pitch and energy contours of speech, have been found to contain speaker-specific information [47]. This prosodic evidence has also been combined with that obtained using spectral features to improve the overall recognition performance. Other prosodic statistics such as mean and variance of pause durations and F_0 values per word have also been used for speaker verification [48]. Methods have also been proposed to use the relation between the dynamics of F_0 and energy trajectories to characterize certain prosodic gestures (rise and fall in the energy or pitch) that are speaker-specific [49]. In addition, these dynamics can also capture the speaking style of the speaker. The extraction and representation of prosodic features from vowel regions has also been proposed [50].

2.1.2.2 Phonetic features

Phonetic features are also referred to as acoustic tokenization features, and capture the spectral characteristics, pronunciation idiosyncrasies, and lexical preferences at the phone level [51]. A phone N-gram approach has been suggested in which a time sequence of phones coming from a bank of open-loop phone recognizers is used to capture speaker-dependent pronunciations [48]. The statistical pronunciation dynamics of phones across multiple languages (cross-stream dimension) has been investigated as an additional component to the time dimension [17]. Speaker detection based on conditional pronunciation modeling has also been researched [52].

2.1.2.3 Lexical features

Lexical features are high-level features which characterize a speaker's distribution of word sequences. In 2001, Doddington proposed a recognition system in which a speaker's characteristic vocabulary (idolect) was used to build a lexical N-gram model [53]. More recently, the approach has been extended to encode the duration of frequent word types as part of the N-gram frequencies [54]. This technique is a hybrid of lexical and prosodic features, since it explicitly models both N-gram frequencies and word durations.

2.2 SPEAKER MODELING

During the enrollment phase of an SIV system, speaker models are constructed using features extracted from the training utterances. In the testing phase, a match score is computed for each feature, which is a measure of how similar the feature vectors extracted from the test utterance are to a particular speaker model. Speaker models can be divided into two categories, namely, template models and stochastic models.

2.2.1 Template Models

In a template model, or non-parametric model, the pattern matching is deterministic in nature. From the 1980s, the usage of different template matching models has been studied for speaker recognition [2, 16]. The template matching method can be dependent or independent of time.

A time-dependent template model must be able to accommodate variability in speaking rate. The most popular method used to compensate for speaking rate variability is the Dynamic Time Warping (DTW) algorithm [4]. This algorithm performs a constrained, piece-wise linear mapping of time axes to align two signals of different lengths. In text-dependent systems, the template model consists of a sequence of feature vectors extracted from a fixed phrase. During the testing phase, DTW is used to align the test features with the template, and then a similarity score is computed.

In a time-independent template model, all temporal variation is ignored and global averages are used. For example, in text-independent systems, a feature averaging method is used, in which the template is the centroid of the set of training features [2]. Another popular time-independent technique, introduced in the 1980s, is Vector Quantization (VQ) [55]. In this method, a codebook is designed for each enrolled speaker by standard clustering procedures, using feature vectors extracted from the training data. The similarity score is given by the distance between an input feature vector and the minimum distance code word in the VQ codebook. The clustering procedure used to form the codebook averages out temporal information, and thus VQ modeling is time-independent in nature.

2.2.2 Stochastic Models

In stochastic models, the pattern matching is probabilistic and results in a measure of the likelihood, or conditional probability, of the observation given the model. By far the most commonly used stochastic modeling technique, introduced in 1995, is the Gaussian Mixture Model (GMM) [7, 56]. A GMM is composed of a weighted mixture of Gaussian components.

The success of GMMs for speaker recognition can be attributed to the fact that a weighted sum of Gaussian probability density functions can be used to model any arbitrary distribution (in this case speaker-dependent feature variations) closely. GMMs have become the state-of-the-art benchmark models in text-independent speaker recognition. A GMM, trained for short-term spectral features, is often taken as the baseline to which new models and features are compared [11]. For this reason, we have selected GMM as the speaker modeling technique in this thesis. A detailed description of speaker modeling using GMM is presented in Chapter 4.

2.2.3 Other Models

In recent times, Support Vector Machines (SVM), which are non-probabilistic binary linear classifiers, have been employed for speaker recognition with low-level as well as high-level

features[46, 57]. SVM has also been successfully combined with GMM to increase accuracy [58]. Artificial neural networks (ANNs) have been utilized in speaker recognition as well [59, 60].

2.3 ROBUST SPEAKER RECOGNITION

As described in the previous chapter, channel mismatch and background noise are major challenges that need to be tackled by real-life speaker recognition systems. While an automatic speaker recognition system performs as well as a human in clean conditions, its performance deteriorates significantly under noise. Noise is very detrimental for SIV in both matched and mismatched conditions, the latter usually being worse. Background noise is usually considered to be additive and non-stationary in nature. Various methods have been proposed to mitigate the effect of channel mismatch and noise on speaker recognition. These methods attempt to improve robustness at the signal level, feature level, model level, or frame level.

2.3.1 Signal Enhancement

At the signal level, a number of speech enhancement techniques and noise compensation techniques have been proposed to suppress the background noise before feature extraction. Such methods include spectral subtraction, minimum mean square error, and combined temporal and spectral processing (CTSP)-based speech enhancement techniques [61]. The disadvantage of enhancement methods is the need for explicit modeling of the noise spectrum, which might pose a challenge when the noise is non-stationary. In addition, signal enhancement methods are computationally very expensive.

2.3.2 Feature Selection and Compensation

Robustness can be improved at the feature level in two ways – robust feature selection and feature compensation. If we select a feature whose properties make it more immune to noise, we can mitigate the effect of noise on the SIV system. Many such robust features were discussed in Section 2.1.

Feature compensation techniques try to counter the effect of background noise during the evaluation phase, by normalizing or adapting the features before scoring [21]. For cepstral features, one method of feature compensation is Cepstral Mean Subtraction (CMS). Since channel effects become additive in the cepstral domain, the mean value of the features over short segments is subtracted from each feature vector before evaluation [2]. However, under additive noise conditions, the feature estimates degrade significantly [62]. As an extension to the above, variance normalization has also been employed in the cepstral domain. A Relative Spectral (RASTA) filtering approach has been proposed, which applies a bandpass filter in the cepstral domain. Various feature transformation methods such as affine transformation, non-linear spectral magnitude normalization, feature warping, and short-time Gaussianization have also been explored [62, 63].

2.3.3 Model Compensation

Model-level compensation involves adapting the speaker model to the environmental conditions instead of the feature vectors. Model compensation may be data-driven, in which the noisy data is used to adapt previously trained speaker models, or analytical, in which noisy speaker models are synthesized from clean speaker models and noise models [21]. An example of a model compensation technique is Speaker Model Synthesis (SMS), in which model parameter changes between different channels are learned and compensated for, allowing synthesis of speaker models in unseen training conditions [64].

Although model compensation techniques improve speaker recognition performance significantly, they usually require prior knowledge of the test environment, which is not possible in real-life scenarios [21].

2.3.4 Frame-level Selection

Instead of modifying features for compensating the effect of noise, features can be extracted from selective regions of speech which are robust against noise. This is known as the missing data (or missing feature) approach. In this approach, a time-frequency analysis of every frame is performed, and the noise in each T-F atom is quantified. Only those T-F atoms that are labeled as speech dominant are used for speaker recognition [64].

A common frame-level selection technique is choosing only those frames which have SNR higher than some threshold for speaker recognition. This is a special case of missing feature theory, that is, all features in a low SNR frame are missed [23]. Different frame-level selection methods are discussed in the following section.

2.4 Relative Importance of Speech Zones

Until now, in the domain of speaker recognition, there has been a lot of focus on feature selection, speaker modeling and robust recognition. However, the relative importance, or the 'speaker-discriminative power' of different regions of speech has not been explored widely. In the following section, we review some of the literature investigating the speaker-discriminative power of various speech zones.

2.4.1 Speaker-discriminative Zones in Speech

In speaker recognition, it is important to select useful zones of the signal, containing speaker specific information, and discard redundant zones of the signal which do not contribute any information. In the training phase, lowering redundancy ensures that the modeling algorithm focuses on the important information, and builds more effective speaker models. Similarly, during the testing phase, more importance should be given to frames containing useful information when compared to redundant frames. An approach known as frame pruning has been examined to reject abnormal frame scores in order to make the system more robust [65]. An information theoretic

approach has been suggested in which feature frames are chosen to have minimum-redundancy within selected feature frames, but maximum-relevancy to speaker models [66].

The first step in frame selection is the removal of non-speech and noise regions. Most speaker recognition systems today only use speech activity detection (SAD), also called voice activity detection (VAD), algorithms to classify every frame as speech or non-speech. Voice activity detection is a very challenging task especially in noisy environments. The most common VAD algorithm is energy-based, in which frames are classified based on a simple energy threshold. However, this algorithm does not work well under background noise. A number of robust, real-time voice activity detection algorithms have been proposed, using various measures such as Periodic to Aperiodic component Ratio (PARADE) [67], Long Term Spectral Divergence (LTSD) [68], Long-term Signal Variability [69], Long-term Spectral Flatness Measure (LSFM) [70], etc.

Even after silence removal, a speech signal contains a lot of redundant information, which can be eliminated during feature extraction. One important factor to consider is that the amount of speaker-specific information varies from one phoneme to another. Studies have shown that among the different phonetic classes, the nasals and vowels are found to be the most speaker discriminative [71]. Phoneme category based speaker recognition has been investigated in a number of papers, and resulted in the finding that vowels provide the best recognition performance [72-74].

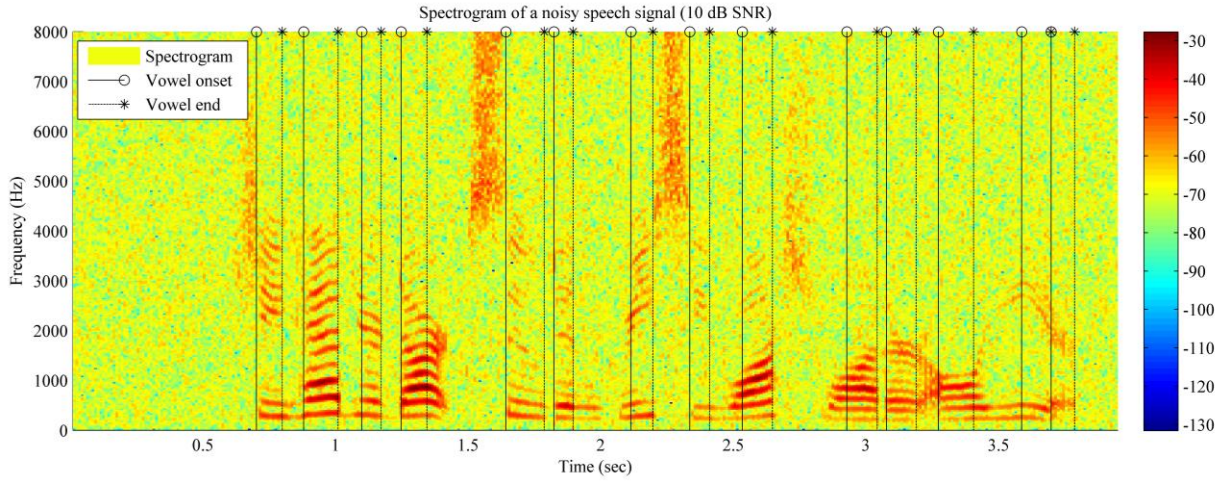

Fig. 2.3: Spectrogram of a noisy speech signal at 10 dB SNR.

Since vowels are produced by an open configuration of the vocal tract, resulting in a relatively unobstructed airflow, they are high-energy regions. In Fig. 2.3, we see that under noisy conditions, vowel regions have a higher SNR, and therefore appear to be very good candidates for feature extraction. In fact, a significant improvement has been observed in the performance of speaker identification and verification under degraded conditions, by using features extracted from vowel and vowel-like regions [75, 76].

Although consonants are extremely important in speech, when they are used on their own, they do not provide good recognition results [66, 74]. This can be attributed to their short duration and low energy. However, when consonants are combined with vowels, i.e. as consonant-vowel (CV) units, they result in very good recognition performance [74]. A similar result has been obtained by Jung et al., confirming that vowels provide the best standalone performance among all phoneme classes [66]. In addition, the lowest overall error rate was in fact achieved by combining consonant classes with vowels, reinforcing the importance of consonants.

Another interesting result was that transition frames, i.e. frames located at phoneme boundaries performed comparably with vowels, even though they were much lower in number [66]. This result might be attributed to the fact that the transition from one sound to another (co-articulation), tends to be speaker-dependent. Further studies on formant transition patterns when a speaker is moving from a consonant to a vowel have confirmed the influence of the speaker's style on CV transitions [77]. Also, for a particular speaker, the formants of different vowels have been observed to portray the same patterns at transition frames [78].

A speech signal can be segmented into quasi steady-state (QSS) zones and transient zones. Quasi steady state zones are of longer duration, and correspond to the middle part of phonemes. The behavior of the speech signal is fairly stationary in these zones. Transient zones, present around phoneme boundaries, are of shorter duration and exhibit rapidly varying spectral characteristics. Upon studying the relative speaker discriminative power of transient and steady zones under clean conditions, it has been found that frames extracted from transient zones are more speaker discriminative than the frames extracted from steady zones of speech [79] [80]. The study showed that the essential information to distinguish between speakers is present in transient zones. Significant computational savings were obtained by discarding steady zones in the testing stage, without much loss in performance. Instead of discarding steady zones completely, a weighted scoring method in which transient zones are given higher weights, was also shown to be effective.

In speaker recognition systems, the speech signal is typically split into 10-30 msec long frames with an overlap of around half the frame length. The assumption made by this method is that the speech signal exhibits quasi-stationary behavior within each frame, and features are extracted from them. However, this assumption might not be valid all the time. Thus, considering a variable frame length could be more effective to extract the varying time-frequency characteristics of the speech signal. In addition, since the steady zones are fairly stationary, we need not repeatedly extract frames from these regions. On the other hand, we need to extract frames more frequently in the dynamically varying transient zones. Therefore, both variable frame length (VFL) as well as variable frame rate (also called frame shift/overlap) (VFR) methods are worth looking into. In 2010, Jung et al. proposed a variable frame length and rate algorithm based on spectral kurtosis for speaker verification [81]. Analyzing the results of the algorithm, it was found that the redundant frames from periodic speech parts such as vowels are relatively reduced. On the other hand, shorter frame lengths and rates were selected in transition regions.

From the studies presented above, it is clear that transient regions of speech encapsulate a lot of speaker-specific characteristics. However, the speaker-discriminative power of these regions under noisy conditions has not been investigated. This is the motivation for our research, in which we study the relative importance of transitions into and out of vowels, i.e. at consonant-vowel boundaries. The reasoning behind choosing only transitions into/out of vowels is that the higher energy content in vowels will facilitate location of these regions of interest under noisy conditions.

2.4.2 Transitions into and out of Vowels

To isolate the transitions into and out of vowels, we propose to first locate the vowel regions in speech, and determine the beginning and end of the vowel regions. Then, we can extract transition frames by examining a small window around every vowel onset and end point. Since vowels are prominent in the signal due to their higher energy and periodic nature, it is easier to locate them compared to other events, especially in noisy conditions. In a typical consonant-vowel-consonant (CVC) syllable, the end of the consonant part and the beginning of the vowel part signifies the Vowel Onset Point (VOP). Similarly, the Vowel End Point (VEP) marks the end of the vowel region.

Recently, there has been a lot of research on the detection of vowel onset points (VOP) and vowel offset/end points (VEP) in speech, under both clean and noisy conditions. Vowel onset and end points can be detected using complementary evidence from a number of sources. In 2009, Prasanna et al. proposed a method in which onset and offset points are detected by combining evidence from the excitation source signal, spectral peaks energy, and modulation spectrum energy [82, 83]. This method has shown to provide very good results under clean conditions, but the number of spurious detections is very high under noise. A two-level method detection algorithm has been proposed [61] , where the onset and offset points identified using the previous method [82] are classified as genuine or spurious, and their locations are corrected using the uniform epoch intervals present in the vowel regions. This method has been shown to perform better under noisy conditions. Recently, Prasanna et al. have suggested methods of detecting vowel like region onset points (VLROPs) using excitation source information [75]. This method can be used for the detection of VOPs as well. The detection of vowel like region end points (VLREPs) was proposed by Pradhan et al. [76]. A method for the detection of VOPs and VEPs using Bessel features has also been explored [84].

The transition into a vowel lies immediately after a VOP, and is around 20-30 milliseconds long. The consonant region lies immediately before the VOP. The duration of most consonants is around 15 to 30 milliseconds, and may be larger for aspirated sounds [85]. Similarly, the transition out of a vowel lies immediately before the VEP, and a consonant region is present immediately after. In the region of 20 milliseconds around a VOP or VEP, the dynamics of the speech signal vary rapidly. In this thesis, we attempt to extract this dynamic information from the speech signal.

3 FEATURE EXTRACTION

In this chapter, we describe the feature extraction module of our speaker identification and verification systems. A block diagram of the feature extraction process is shown in Fig. 3.1. The process of feature extraction consists of six major steps, namely:

1. Pre-emphasis
2. Frame Blocking
3. Windowing
4. Voice Activity Detection
5. Linear Prediction
6. Conversion to Line Spectral Frequencies

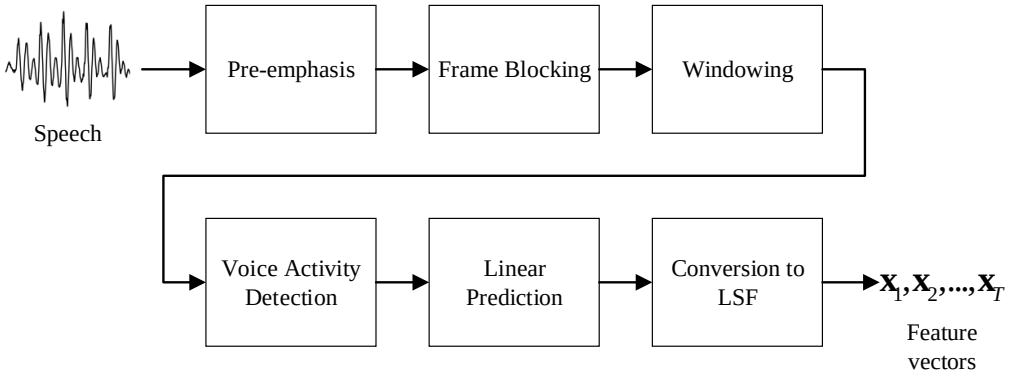

Fig. 3.1: Feature extraction process.

Before delving into a description of the feature extraction process, let us look at a mathematical representation of the speech production process. The source-filter model of speech production, depicted in Fig. 3.2, was proposed by Gunnar Fant in the 1950s [86].

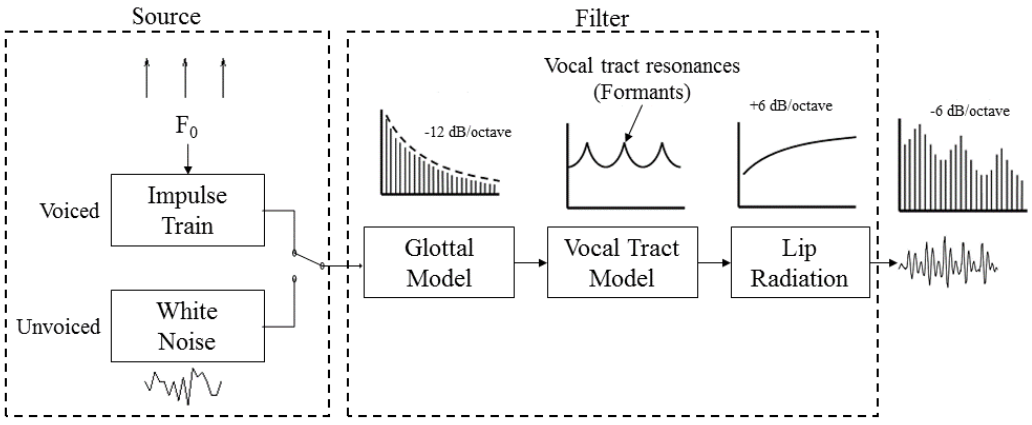

Fig. 3.2: Source-filter model of speech production.

According to the source-filter model, speech is the response of a filter, i.e. the vocal tract, to a source, i.e. an excitation signal from the glottis. During voiced speech, the vocal cords vibrate, causing the glottis to open and close rapidly. Thus, the excitation source is a pulsating air stream passing through the glottis, and is mathematically represented by an impulse train. Unvoiced speech is produced by turbulent airflow due to a constriction in the vocal tract. In this case, the excitation signal is represented by white noise. In the linear prediction (LP) analysis of speech, we calculate the parameters of the vocal tract filter that transformed the excitation signal into a speech signal.

In reality, however, the spectrum of the glottal source signal rolls off at approximately -12 dB/octave [87]. This roll-off is accounted for by the glottal model block in Fig. 3.2. When speech is radiated from the lips, low frequency components are attenuated, causing a spectral rise of around 6 dB/octave. This radiation characteristic is modeled by the lip radiation block. The spectrum of a speech signal rolls-off at approximately -6 dB/octave, as a result of the -12 dB/octave slope of the glottal source and the 6 dB/octave spectral rise due to lip radiation. Thus, the filter calculated by LP analysis also includes the effects of the glottal source shape and lip radiation.

3.1 PRE-EMPHASIS

Since we want to model the vocal tract response, we need to counter the spectral roll-off of speech before LP analysis. This is done by passing the speech signal through a first order high-pass filter, also known as a pre-emphasis filter. This boosts the high frequency content and compensates for the spectral roll-off.

Let us assume that a continuous-time speech signal $s(t)$ was recorded at a sampling frequency of f_s Hz. Then, the finite length, discrete-time speech signal of length N is denoted by:

$$s[n] = s(nT_s), \quad T_s = \frac{1}{f_s}, \; 0 \le n \le N-1 \tag{3.1}$$

We use a pre-emphasis filter whose transfer function is:

$$H_p(z) = \frac{Y(z)}{S(z)} = 1 - \alpha z^{-1} \tag{3.2}$$

where $Y(z)$ and $S(z)$ are the z-transforms of the pre-emphasized signal $y[n]$ and the speech signal $s[n]$. In the time domain, the output signal after pre-emphasis is given by:

$$y[n] = s[n] - \alpha s[n-1] \tag{3.3}$$

Typically, a value of α in the range [0.9, 1] is chosen. In our research, we have chosen α = 0.97, following previous speaker identification systems [64]. The effect of pre-emphasis on a speech

signal is illustrated in Fig. 3.3. It can be clearly seen that the high frequency content has been emphasized after filtering.

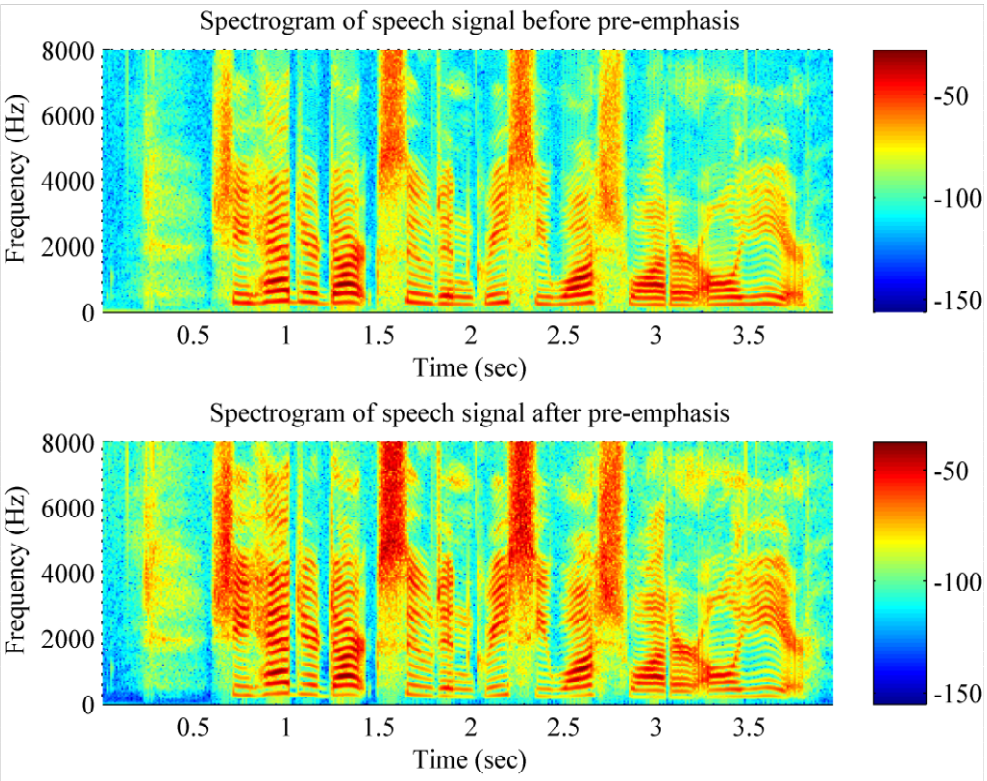

Fig. 3.3: Effect of pre-emphasis on a speech signal.

3.2 FRAME BLOCKING

Next, the pre-emphasized signal is divided, or 'blocked', into overlapping frames of fixed length, as illustrated in Fig. 3.4. Overlapping is performed to ensure temporally smoother parameter transitions between frames. Let the frame length be L samples, and the frame shift be δ samples. A speech signal of length N can be divided into Γ frames, where

$$\Gamma = \left\lfloor \frac{N-L}{\delta} \right\rfloor + 1 \tag{3.4}$$

The n^{th} sample in the t^{th} frame is denoted by:

$$f_t[n] = \begin{cases} y[(t-1)\delta + n], & 0 \le n \le L-1 \\ 0, & otherwise \end{cases}, \quad 1 \le t \le \Gamma \tag{3.5}$$

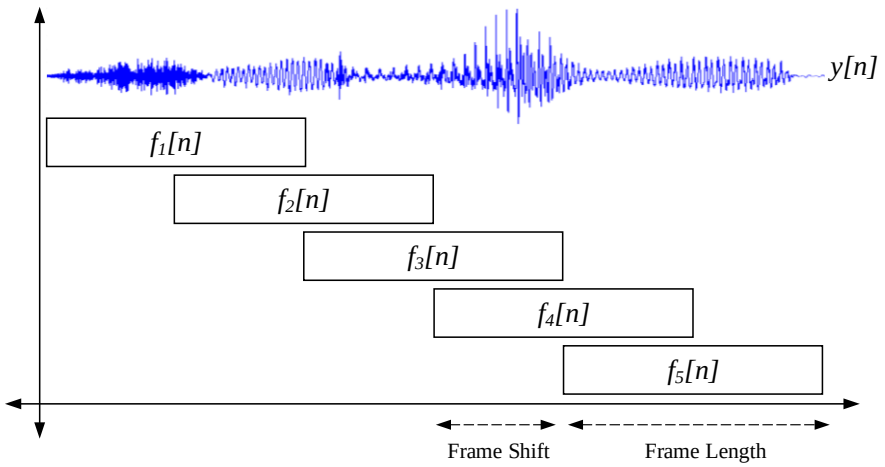

Fig. 3.4: Frame blocking.

3.3 WINDOWING

After the speech signal has been divided into frames, each frame is multiplied with a Hamming window function. A Hamming window of length L is given by:

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{L-1}\right), & 0 \le n \le L-1 \\ 0, \quad otherwise \end{cases} \tag{3.6}$$

Then, the t^{th} frame after windowing, is given by:

$$x_t[n] = \begin{cases} f_t[n]w[n], & 0 \le n \le L-1 \\ 0, \quad otherwise \end{cases}, \quad 1 \le t \le \Gamma \tag{3.7}$$

Windowing is performed to taper each frame at the ends, and reduce discontinuities at the frame edges. When no windowing is performed, it is as if the frame was multiplied with a rectangular window. The edge discontinuities that occur in the case of a rectangular window affect the estimation of linear prediction coefficients, which in turn affects the Line Spectral Frequency values. This is explained in detail in Section 3.5. The effect on windowing on the Line Spectral Frequency coefficients of a speech signal is shown in Fig. 3.5.

Figure 3.5 shows the Line Spectral Frequencies of order 8, obtained using 20 msec long frames, with 10 msec frame shift. We observe smooth variations in the LSFs when a Hamming window is used. In contrast, a rectangular window results in heavy oscillations in the LSF estimation [88]. These oscillations are very detrimental for speaker identification, especially when spectro-temporal features are extracted.

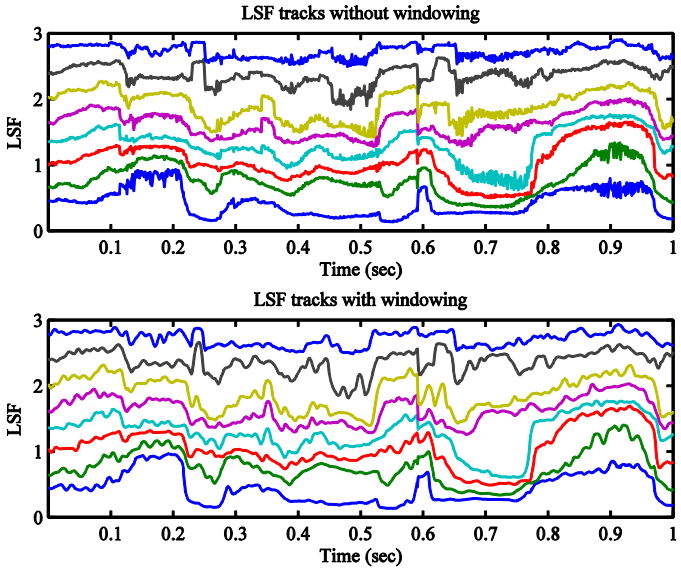

Fig. 3.5: A comparison of LSF tracks with and without windowing.

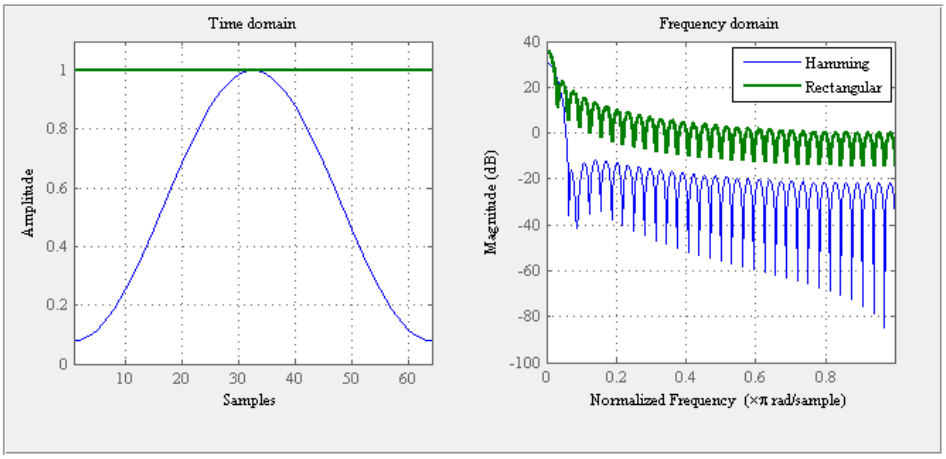

Fig. 3.6: A comparison of Rectangular window versus Hamming window.

Next, let us look at the effect of widowing in the frequency domain. The multiplication in the time domain translates to a convolution in the frequency domain. From Fig. 3.6, we see that in the frequency domain, the side lobes of a rectangular window are much higher than those of a Hamming window. The high side lobes of the rectangular window result in spectral leakage, causing undesirable spectral artefacts.

On the other hand, the Hamming window has a very high energy concentration in the main lobe when compared to the side lobes. This helps to minimize spectral leakage, creating a smoother and less distorted spectrum. Since we are ultimately interested in extracting LSFs, which are short-term spectral features, we wish to minimize these spectral artifacts. Therefore, windowing is a crucial pre-processing step in our speaker identification system.

3.4 VOICE ACTIVITY DETECTION

A simple energy-based voice activity detector [2, 89] is used to classify each frame as speech or non-speech. The voice activity detection is based on the assumption that low and high energy frames, respectively, correspond to non-speech and speech. First, the log-energy of the t^{th} frame is computed as:

$$E_t = 10 log_{10}\left(\frac{1}{L-1}\sum_{n=0}^{L-1}\left(x_t[n]-\mu_t\right)^2+\epsilon\right), \quad 1 \le t \le \Gamma \tag{3.8}$$

The arbitrary constant $\epsilon = 10^{-16}$ is used to avoid log of zero, and μ_t is the sample mean of the t^{th} frame, given by:

$$\mu_t = \frac{1}{L}\sum_{n=0}^{L-1} x_t[n], \quad 1 \le t \le \Gamma \tag{3.9}$$

Next, we find the maximum energy over all T frames of the utterance:

$$E_{max} = max_{t=1,..,\Gamma}\{E_t\} \tag{3.10}$$

The VAD decision threshold is adjusted according to this maximum energy level. Additionally, a minimum energy threshold is used to avoid false positives. Thus, the condition for a frame to be classified as a speech frame is:

$$\left(E_t > E_{max} - \tau_1\right) \wedge \left(E_t > \tau_2\right) \tag{3.11}$$

τ_1=30 dB and τ_2 =-55 dB are chosen and known as the primary and minimum energy thresholds [89]. The output of the VAD algorithm on a speech signal is shown in Fig. 3.7.

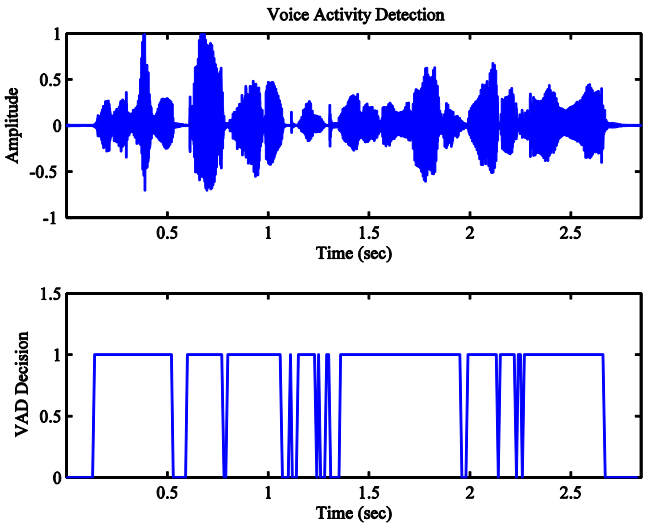

Fig. 3.7: Output of the voice activity detection algorithm.

Let us say that among the Γ frames, *T* frames were classified as speech and $(\Gamma - T)$ were classified as non-speech. Only the features extracted from these *T* speech frames are used for training the speaker models, as well as testing the speaker identification system.

3.5 LINEAR PREDICTION

According to the source-filter model of speech production, a speech wave can be thought of as the response of the vocal tract (filter) to an excitation signal arising from the glottis (source). The excitation signal is a periodic glottal pulse in the case of voiced speech, and is similar to white noise in the case of unvoiced speech. The vocal tract is modeled as a linear, slowly varying filter. An important assumption made by the source-filter model is that the vocal tract and the excitation are independent of each other.

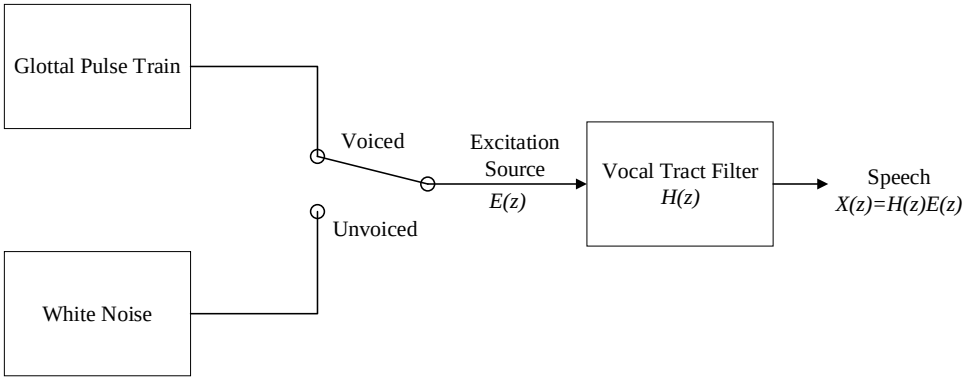

Fig. 3.8: Simplified source-filter model of speech production.

Continuing our assumption that the speech signal is stationary within a frame, we can say that the vocal tract is time-invariant in that frame. In the linear prediction analysis of speech, we consider the vocal tract to be an all-pole filter of order *p* in which the poles correspond to the formant frequencies of the vocal tract. The transfer function of the vocal tract within a frame is written as:

$$H(z) = \frac{1}{A_p(z)} \tag{3.12}$$

$$A_p(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{3.13}$$

$H(z)$ is called the synthesis filter, and $A_p(z)$ is called the inverse filter. The roots of $A_p(z)$ correspond to the resonant frequencies of the vocal tract, i.e. the formant frequencies.

Now, let $x[n]$ be a speech frame of length L and let $X(z)$ be its z-transform. Let $e[n]$ be the corresponding excitation signal, and $E(z)$ its z-transform. Then, we can write

$$X(z) = H(z)E(z) \tag{3.14}$$

Substituting (3.13) into the above, we get

$$X(z) = E(z)\frac{1}{1+\sum_{k=1}^{p} a_k z^{-k}} \tag{3.15}$$

In the time domain, this corresponds to:

$$x[n] = -\sum_{k=1}^{p} a_k x[n-k] + e[n] \tag{3.16}$$

From the above, we can see that the speech sample at instant *n* can be estimated as a linear combination of the previous *p* samples, as

$$\hat{x}[n] = -\sum_{k=1}^{p} a_k x[n-k] \tag{3.17}$$

This is known as *linear prediction*. Then, the excitation signal $e[n]$ can also be thought of as the linear prediction error or residual, i.e.

$$e[n] = x[n] - \hat{x}[n] \tag{3.18}$$

Since $x[n]$ is non-zero only for $0 \le n \le L-1$, $e[n]$ is non-zero only for $0 \le n \le L-1+p$. Now, since the speech signal $x[n]$ is highly correlated in nature, we can solve for the LP filter coefficients by minimizing the total squared prediction error, given by:

$$E = \sum_{n=-\infty}^{\infty} \left(e[n]\right)^2 \tag{3.19}$$

Thus, using (3.17) and (3.18), we get

$$E = \sum_{n=0}^{L-1+p} \left(x[n] + \sum_{k=1}^{p} a_k x[n-k] \right)^2 \tag{3.20}$$

For values of *n* less than *p* we are predicting the signal from zero-valued samples outside the frame range. As a result, $e[n]$ will be relatively large. Similarly, at values of n greater than *L*, we are predicting zero valued samples outside the frame range, from non-zero samples. Thus, $e[n]$ will again be relatively large. To reduce the effect of these meaningless edge components on the quantity to be minimized, a tapering window function, such as a Hamming window is used before linear prediction [78].

The minimum of *E* can be found by differentiating (3.20) with respect to every a_j, and setting those derivatives to zero, i.e.

$$\frac{\partial E}{\partial a_j} = 0, \quad 1 \le j \le p \tag{3.21}$$

Substituting (3.20) into (3.21), we get

$$\sum_{n=0}^{L-1+p} 2\left(x[n] + \sum_{k=1}^{p} a_k x[n-k] \right)\left(x[n-j] \right) = 0, \quad 1 \le j \le p \tag{3.22}$$

$$\sum_{k=1}^{p} a_k \sum_{n=0}^{L-1+p} x[n-k]x[n-j] = -\sum_{n=0}^{L-1+p} x[n]x[n-j], \quad 1 \le j \le p \tag{3.23}$$

Substituting $m = n - j$ in the summations, and retaining only non-zero products, the above equations can be further reduced to:

$$\sum_{k=1}^{p} a_k \sum_{m=0}^{L-1+k-j} x[m+j-k]x[m] = -\sum_{m=0}^{L-1-j} x[m+j]x[m], \quad 1 \le j \le p \tag{3.24}$$

The set of equations obtained above can be solved using the *autocorrelation method* [90, 91]. Let us consider the autocorrelation of $x[n]$, given by:

$$\begin{aligned} R_x(j) &= \sum_{m=-\infty}^{\infty} x[m]x[m+j] \\ &= \sum_{m=0}^{L-1-j} x[m]x[m+j] \end{aligned} \tag{3.25}$$

Then, we can show that:

$$\begin{aligned} R_x(j-k) &= \sum_{m=-\infty}^{\infty} x[m]x[m+j-k] \\ &= \sum_{m=0}^{L-1-j+k} x[m]x[m+j-k] \end{aligned} \tag{3.26}$$

Substituting (3.25) and (3.26) in (3.24), we get,

$$\sum_{k=1}^{p} a_k R_x(j-k) = -R_x(j), \quad 1 \le j \le p \tag{3.27}$$

The set of equations obtained above are known as the Yule-Walker equations, and can be written in matrix form as:

$$\begin{bmatrix} R_x(0) & R_x(1) & \cdots & R_x(p-1) \\ R_x(1) & R_x(0) & \cdots & R_x(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_x(p-1) & R_x(p-2) & \cdots & R_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_x(1) \\ R_x(2) \\ \vdots \\ R_x(p) \end{bmatrix} \tag{3.28}$$

The square matrix obtained above is a Toeplitz matrix. It is symmetric, with all elements on each diagonal being identical, and is easy to invert [90]. The system of equations in (3.28) can be solved using the Levinson-Durbin recursion algorithm, described below, in $O(p^2)$ time [91]:

$$E_0 = R_x(0) \tag{3.29}$$

$$k_i = -\frac{\left\{R_x(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R_x(i-j)\right\}}{E_{i-1}} \tag{3.30}$$

$$a_i^{(i)} = k_i \tag{3.31}$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \tag{3.32}$$

$$E_i = \left(1 - k_i^2\right) E_{i-1} \tag{3.33}$$

The above equations (3.29)-(3.33) are solved recursively for $i = 1, 2, \ldots, p$ to obtain the linear prediction coefficients:

$$a_j = a_j^{(p)}, \quad 1 \le j \le p \tag{3.34}$$

3.6 CONVERSION TO LSF

Once the linear prediction coefficients have been estimated for each frame, they are converted to a Line Spectral Frequency (LSF) representation. According to the acoustic tube model of speech production, the vocal tract can be viewed as a tube of varying diameter. The resonant frequencies of the tube correspond to the formant frequencies. During voiced speech, the glottis opens and closes rapidly, i.e. it is neither fully open nor fully closed. Thus, speech production can be considered a combination of two resonance conditions – vocal tract tube closed at the glottis end and vocal tract tube open at the glottis end [30].

As described in the earlier section, the inverse filter polynomial obtained by p^{th} order LP analysis is given by:

$$A_p(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{3.35}$$

Assuming that $A_p(z)$ arises from a combination of two resonance conditions, it can be written as a combination of two polynomials $P(z)$ and $Q(z)$, both of order *(p+1)*.

$$A_p(z) = \frac{P(z) + Q(z)}{2} \tag{3.36}$$

P(*z*) is a symmetric polynomial which describes the resonance conditions arising due to complete closure of the vocal tract at the glottis. The feedback term is positive to model energy reflection at a completely closed glottis.

$$\begin{aligned} P(z) &= A_p(z) + z^{-(p+1)} A_p\left(z^{-1}\right) \\ &= 1 + \sum_{k=1}^{p} \left(a_k + a_{p+1-k}\right) z^{-k} + z^{-(p+1)} \end{aligned} \tag{3.37}$$

On the other hand, *Q*(*z*) is an anti-symmetric polynomial which describes the resonance conditions arising due to the complete opening of the glottis. The feedback term is negative to model energy reflection at a completely open glottis.

$$\begin{aligned} Q(z) &= A_p(z) - z^{-(p+1)} A_p\left(z^{-1}\right) \\ &= 1 + \sum_{k=1}^{p} \left(a_k - a_{p+1-k}\right) z^{-k} - z^{-(p+1)} \end{aligned} \tag{3.38}$$

The Line Spectral Frequencies are the angular positions (phases) of the roots of *P*(*z*) and *Q*(*z*) The *p+1* roots of *P*(*z*) can be written as:

$$\theta_{Pk} = e^{j\omega_{Pk}}, \quad 1 \le k \le p+1 \tag{3.39}$$

Then, the phases of the roots of *P*(*z*) are given by:

$$\omega_{Pk} = \tan^{-1}\left(\frac{Re\{\theta_{Pk}\}}{Im\{\theta_{Pk}\}}\right), \quad 1 \le k \le p+1 \tag{3.40}$$

Similarly the *p+1* roots of *Q*(*z*) can be written as

$$\theta_{Qk} = e^{j\omega_{Qk}}, \quad 1 \le k \le p+1 \tag{3.41}$$

Then, the phases of the roots of *Q*(*z*) are given by:

$$\omega_{Qk} = \tan^{-1}\left(\frac{Re\{\theta_{Qk}\}}{Im\{\theta_{Qk}\}}\right), \quad 1 \le k \le p+1 \tag{3.42}$$

Four interesting observations can be made from the above:

1. The roots of $P(z)$ and $Q(z)$ lie on the unit circle in the complex plane.
2. If p is even, –1 is a root of $P(z)$ while 1 is a root of $Q(z)$. If p is odd, –1 and 1 are both roots of $Q(z)$ [92].
3. Since the coefficients of $P(z)$ and $Q(z)$ are real, the rest of the $2p$ complex roots occur in conjugate pairs. Thus, the angular positions of only p roots (conventionally between 0 to π) need to be stored.
4. The roots of $P(z)$ are interlaced with those of $Q(z)$. If p is even, we can write:

$$0 < \omega_{P1} < \omega_{Q1} < \omega_{P2} < \omega_{Q2} \ldots . < \omega_{\frac{Pp}{2}} < \omega_{\frac{Qp}{2}} < \pi \tag{3.43}$$

Thus, the Line Spectral Frequency (LSF) feature vector of the frame is chosen as:

$$\mathbf{x} = \left[\omega_{P1}\, \omega_{Q1}\, \omega_{P2}\, \omega_{Q2} \ldots \omega_{\frac{Pp}{2}}\, \omega_{\frac{Qp}{2}} \right] \tag{3.44}$$

In order to visualize the LSF representation, a 20 msec long speech frame is considered. LP analysis of order 20 is performed. The poles of $H(z)$ (or zeros of $A_p(z)$), and the zeros of $P(z)$ and $Q(z)$ are shown in Fig. 3.9. We see that the roots of $P(z)$ and $Q(z)$ are interlaced on the unit circle. The LSFs are the angular positions of these roots (excluding roots at -1 and 1). If a zero of $A_p(z)$ lies close to the unit circle, its angular position is bracketed by a pair of LSFs. For the zeros of $A_p(z)$ on the real axis, the bracketing property is not observed.

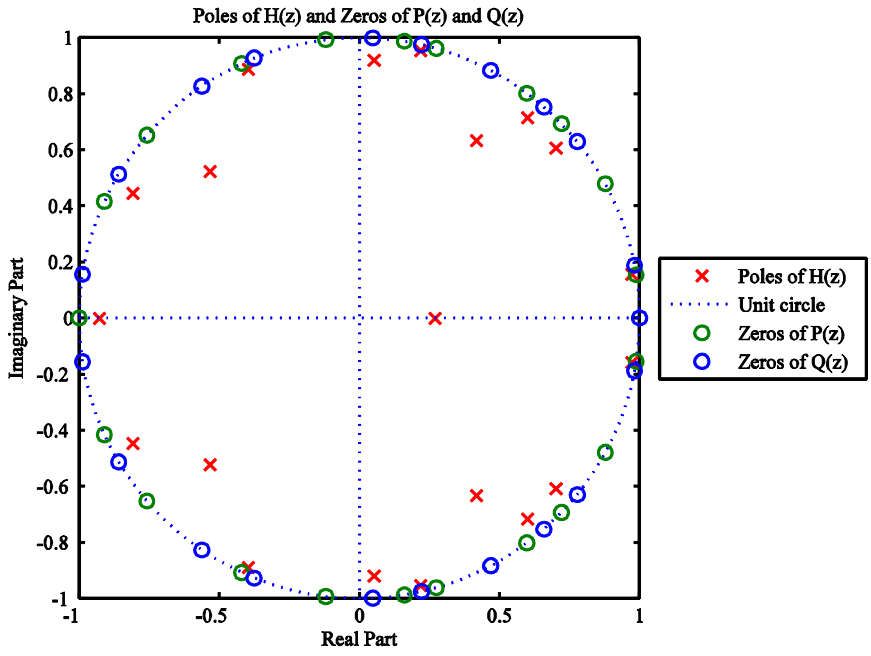

Fig. 3.9: LP Poles and zeros of the LSF polynomials

The FFT spectrum of the frame is shown in Fig. 3.10. The magnitude response of the all-pole filter obtained, and the line spectral frequencies are overlaid on the speech spectrum. From the figure, we see that the LP spectrum approximates the speech spectrum envelope. Another important observation is that the Line Spectral Frequencies represent a bracketing of the formants. Consider the magnitude response of the vocal tract filter, given by:

$$\left|H\left(e^{j\omega}\right)\right| = \frac{1}{\left|A_p\left(e^{j\omega}\right)\right|} = \frac{2}{\left|P\left(e^{j\omega}\right)+Q\left(e^{j\omega}\right)\right|} \tag{3.45}$$

LSFs are the phases of the zeros of $P(z)$ and $Q(z)$. Hence, if a pair of LSFs are very close to ω_0, then $\left|P\left(e^{j\omega_0}\right)+Q\left(e^{j\omega_0}\right)\right|$ will be very close to zero, resulting in a peak around ω_0 in the magnitude response curve [93]. Thus, as shown in Fig. 3.10, every formant peak is bracketed by a pair of LSFs. On the contrary, if a pair of LSFs are far from each other, the magnitude response curve will be relatively flat around the two LSFs.

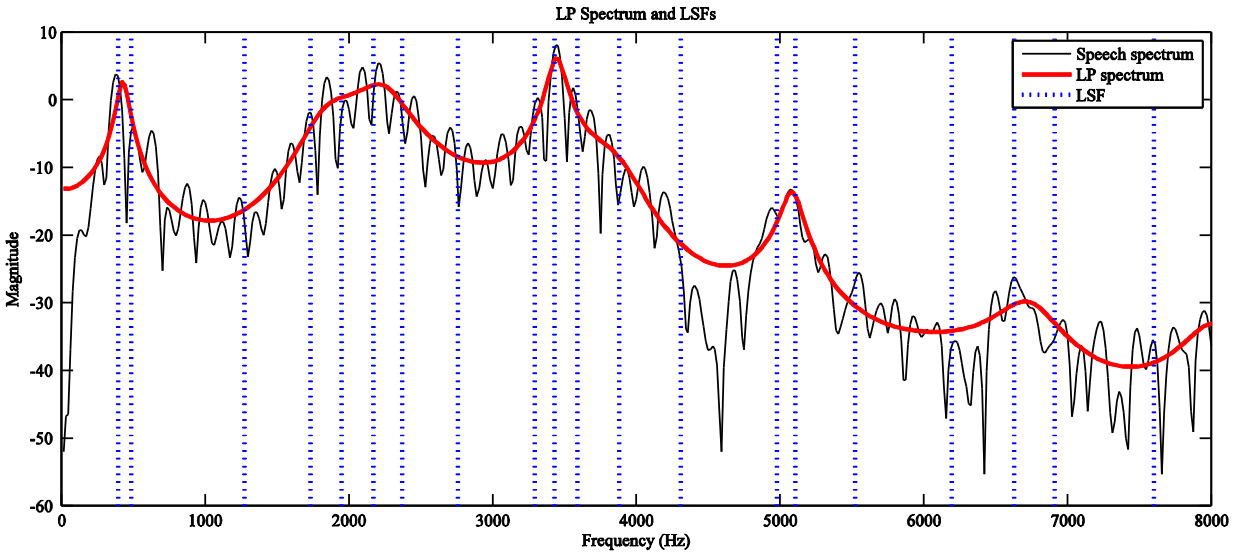

Fig. 3.10: LP spectrum and Line Spectral Frequencies.

4 SPEAKER IDENTIFICATION AND VERIFICATION

An automatic speaker identification (SI) or speaker verification (SV) system is developed and evaluated in two phases: the enrollment/training phase and the testing (identification/verification) phase. This chapter outlines the speaker modeling techniques used in the enrollment phase, and also describes the testing phase of our SI and SV systems.

4.1 SPEECH AND NOISE CORPORA

In this section, the speech and noise corpora used for training and testing our speaker identification and verification systems are described.

4.1.1 TIMIT Speech Corpus

In our research, the TIMIT speech corpus is used for training and testing the speaker identification and verification systems. The TIMIT (Texas Instruments-Massachusetts Institute of Technology) speech corpus is an acoustic-phonetic speech database specifically designed for training and testing automatic speech recognition systems [94]. It is now being widely used in speaker recognition systems as well.

The TIMIT speech corpus contains broadband recordings of 630 speakers, 70% male and 30% female, from eight major dialect regions in America. The dialect distribution of the speakers is shown in Table 4.1.

Table 4.1: Speaker distribution in the TIMIT database by dialect.

Dialect Region Code	Dialect Region	# Male Speakers	# Female Speakers	Total
DR1	New England	31	18	49
DR2	Northern	71	31	102
DR3	North Midland	79	23	102
DR4	South Midland	69	31	100
DR5	Southern	62	36	98
DR6	New York City	30	16	46
DR7	Western	74	26	100
DR8	Army Brat	22	11	33
Total		**438**	**192**	**630**

Each speaker has been recorded reading ten phonetically rich sentences in English, and each sentence is stored as a 16-bit, 16 kHz speech waveform file. Each sentence is approximately 3 seconds long. The speech was recorded in a single session, using a high quality microphone in a sound proof booth. The database also includes time-aligned orthographic, phonetic, and word transcriptions of the speech.

The speech material for each speaker consists of:

1. 2 dialect "shibboleth" (SA) sentences, meant to expose the dialectal variants of the speakers and common to all 630 speakers.
2. 5 phonetically compact (SX) sentences, designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest.
3. 3 phonetically diverse (SI) sentences, selected from existing text sources so as to add diversity in sentence types and phonetic contexts.

The TIMIT corpus has been divided into a TRAIN directory, consisting of 462 speakers, and a TEST directory, containing 168 speakers. The dialect distribution of speakers in the TEST directory is shown in the following table.

Table 4.2: Speaker distribution in the TEST directory of the TIMIT corpus.

Dialect Region Code	Dialect Region	# Male Speakers	# Female Speakers	Total
DR1	New England	7	4	11
DR2	Northern	18	8	26
DR3	North Midland	23	3	26
DR4	South Midland	16	16	32
DR5	Southern	17	11	28
DR6	New York City	8	3	11
DR7	Western	15	8	23
DR8	Army Brat	8	3	11
Total		**112**	**56**	**168**

4.1.2 SPIB Noise Database

The noise signals used in our research are obtained from the Signal Processing Information Base (SPIB) noise database [95]. This dataset contains noise data measured in the field by the Speech Research Unit (SRU) at the Institute for Perception-TNO, the Netherlands.

All noise files have a duration of 235 seconds and are acquired by considering a sampling rate of 19.98 kHz, an analog to digital converter (A/D) with 16 bits, an anti-aliasing filter, and without a pre-emphasis stage. The corpus contains fifteen different types of noise data, from which we selected six noise types, described in Table 4.3. These noise files were downsampled to 16 kHz for our evaluation.

Table 4.3: Noise categories selected from the SPIB dataset.

Noise Filename	Description	Type
White Noise	Acquired by sampling a high-quality analog noise generator (Wandel & Goltermann), which results in equal energy per Hz bandwidth.	White
Pink Noise	Acquired by sampling a high-quality analog noise generator (Wandel & Goltermann), yielding equal energy per 1/3 octave.	Pink
Speech Babble	Acquired by recording samples from 1/2" B&K condensor microphone onto digital audio tape (DAT). The source of this babble is 100 people speaking in a canteen. The room radius is over two meters; therefore, individual voices are slightly audible. The sound level during the recording process was 88 dBA.	Babble
Factory Floor Noise 1	Acquired by recording samples from 1/2" B&K condensor microphone onto digital audio tape (DAT). This noise was recorded near plate-cutting and electrical welding equipment.	Factory
Cockpit Noise 3 (F-16)	Acquired by recording samples from 1/2" B&K condensor microphone onto digital audio tape (DAT). The noise was recorded at the co-pilot's seat in a two-seat F-16, traveling at a speed of 500 knots, and an altitude of 300-600 feet. The sound level during the recording process was 103 dBA.	Cockpit
Vehicle Interior Noise (Volvo 340)	Acquired by recording samples from 1/2" B&K condensor microphone onto digital audio tape (DAT). This recording was made at 120 km/h, in 4th gear, on an asphalt road, in rainy conditions.	Car

4.2 SPEAKER IDENTIFICATION SYSTEM

In this section, we describe the enrollment and testing (or identification) phase of our speaker identification (SI) system.

4.2.1 Training and Test Sets

All SI experiments are conducted using the 168 speakers in the TEST directory of the TIMIT corpus. The speech material of each speaker in the TEST directory is divided into mutually exclusive training and test sets. The training set consists of the 5 SX sentences and the 3 SI sentences. During the enrollment phase, speech from the training set is used to develop a speaker model for each speaker. The test set consists of the two remaining SA sentences. Since we wish to build a text-independent SI system, the two SA sentences that are common to all speakers are not used in training. During the testing phase, speech from the test set is used to conduct speaker identification tests and evaluate the performance of the system.

4.2.2 Speaker Enrollment Using GMM

During the enrollment or training phase, a speaker model is constructed for each speaker using the training speech described above. The feature extraction module described in the previous chapter is used to extract a set of feature vectors from the training speech.

Let us suppose that we want to build a speaker database of S speakers. Assuming that the training speech for the s^{th} speaker consisted of T speech frames, each of which yielded a D-dimensional feature vector. Then, the set of training feature vectors for the s^{th} speaker is denoted by:

$$X_s = \left\{ \mathbf{x}_t \in \mathbb{R}^D : 1 \le t \le T \right\} \tag{4.1}$$

Now, using these feature vectors we need to build a statistical model λ_s for this speaker. In our SI system, we use simple Gaussian Mixture Models (GMM) for the purpose of speaker modeling. The enrollment process is shown in Fig. 4.1.

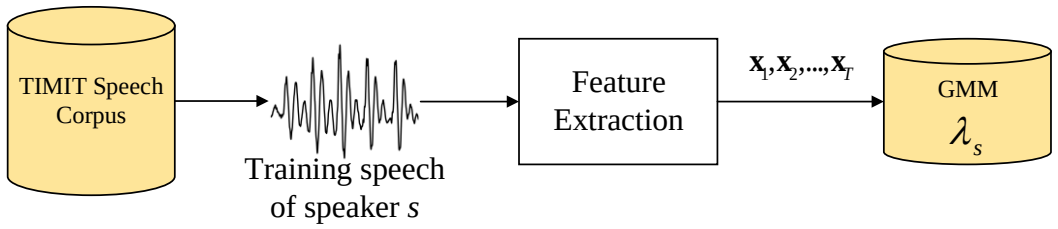

Fig. 4.1: Speaker enrollment process.

4.2.2.1 Gaussian Mixture Models

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a linear weighted sum of Gaussian component densities. The reasons for using Gaussian Mixture Models for speaker identification are two-fold.

First, we hypothesize that the acoustic space of spectral features consists of acoustic classes corresponding to broad phonetic events such as vowels, nasals, or fricatives [7]. The acoustic classes represent certain speaker-dependent vocal tract configurations that are useful for characterizing speaker identity. Then, the individual Gaussian component densities in a GMM might model some of these hidden acoustic classes. Since English speech has no more than 45 or so distinct phones, if the number of components is made suitably large, we might even be able to model all of the phonetic classes [64]. Furthermore, the linear weighing of the different components in a GMM enables smooth transitions from one acoustic class to another, making the system text-independent in nature.

Another reason for using GMM for speaker identification stems from the power of a Gaussian distribution. A linear combination of Gaussian distributions is capable of forming smooth approximations to arbitrarily shaped densities [96].

A Gaussian Mixture Model λ is a weighted sum of M component densities. It is characterized by the probability density function:

$$p(\mathbf{x}|\lambda) = \sum_{m=1}^{M} p_m g_m(\mathbf{x} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{4.2}$$

Here, $\mathbf{x}$ is a D dimensional feature vector, and $g_m(\mathbf{x} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), m = 1, 2, ..., M$ are the individual component densities and $p_m, m = 1, 2, ..., M$ are the mixture weights.

Each component density is a D-variate Gaussian function of the form:

$$g_m(\mathbf{x} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_m|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\mathbf{x} - \boldsymbol{\mu}_m)\right\} \tag{4.3}$$

with mean vector $\boldsymbol{\mu}_m \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma}_m \in \mathbb{R}^{D \times D}$. In addition, the mixture weights satisfy the following constraint:

$$\sum_{m=1}^{M} p_m = 1 \tag{4.4}$$

A Gaussian Mixture Model (GMM) λ is parametrized by the mean vectors, covariance matrices and mixture weights from all M component densities.

$$\lambda = \{p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}, \quad m = 1, 2, \ldots, M \tag{4.5}$$

The GMM can take on several forms based on the type of covariance matrices used. The covariance matrices may be full or diagonal. In addition, they could also satisfy one of the following conditions:

1. *Nodal Covariance*: The model has one covariance matrix for every Gaussian component.
2. *Grand Covariance*: The model has the same covariance matrix for all Gaussian components in one speaker model.
3. *Global Covariance*: The model has the same covariance matrix for all Gaussian components in all speaker models.

In our speaker identification system, we have used *nodal*, *diagonal* covariance matrices. This choice is based on previous research indicating better speaker identification performance when nodal, diagonal covariance matrices are used when compared to using nodal and grand full covariance matrices [7]. Even if the features are not statistically independent, full covariance matrices are not necessary. This is because a linear combination of Gaussian densities with diagonal covariance matrices is also capable of modeling the correlations between feature vector elements. The effect of using a set of full covariance Gaussians can be equally obtained by using a larger set of diagonal covariance Gaussians [64]. In addition, the use of diagonal covariance matrices is also more computationally efficient.

4.2.2.2 Gaussian Mixture Model Training

Given a set of training features, $X = \left\{ \mathbf{x}_t \in \mathbb{R}^D : 1 \le t \le T \right\}$, and a GMM configuration (number of components and type of covariance matrices), we need to find the parameters of a GMM λ that best approximates the distribution of the training features. The most popular method used to determine these parameters is Maximum Likelihood (ML) estimation.

The aim of ML estimation is to determine the values of $\left\{ p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m \right\}$, $m = 1, 2, \ldots, M$, which maximize the likelihood of the GMM λ, given a set of training features X. Assuming independence of the training features X, the GMM likelihood can be written as:

$$p\left(X \mid \lambda\right) = \prod_{t=1}^{T} p\left(\mathbf{x}_t \mid \lambda\right) \tag{4.6}$$

The above function is non-linear in nature, and thus, direct maximization is not possible. Instead, the parameter estimates can be obtained iteratively using a special case of the expectation-maximization (EM) algorithm [96].

In the i^{th} iteration, the EM algorithm begins with a GMM, $\lambda^{(i)} = \left\{ p_m^{(i)}, \boldsymbol{\mu}_m^{(i)}, \boldsymbol{\Sigma}_m^{(i)} \right\}$, $m = 1, 2, \ldots, M$ and estimates a new GMM, $\lambda^{(i+1)} = \left\{ p_m^{(i+1)}, \boldsymbol{\mu}_m^{(i+1)}, \boldsymbol{\Sigma}_m^{(i+1)} \right\}$, $m = 1, 2, \ldots, M$, such that:

$$p\left(X \mid \lambda^{(i+1)}\right) \ge p\left(X \mid \lambda^{(i)}\right) \tag{4.7}$$

Every iteration consists of an expectation step (E-step) and a maximization step (M-step).

E-step: For each feature vector $\mathbf{x}_t$, we compute the *a posteriori* probabilities of this feature vector belonging to each of the *M* Gaussian components in the initial GMM, $\lambda^{(i)}$. These probabilities are also known as the *membership weights*.

$$\begin{aligned} \gamma_{t,m} &= p\left(m \mid \mathbf{x}_t, \lambda^{(i)}\right) \\ &= \frac{p_m g_m\left(\mathbf{x}_t \mid \boldsymbol{\mu}_m^{(i)}, \boldsymbol{\Sigma}_m^{(i)}\right)}{\sum_{k=1}^{M} p_k g_k\left(\mathbf{x}_t \mid \boldsymbol{\mu}_k^{(i)}, \boldsymbol{\Sigma}_k^{(i)}\right)}, \quad 1 \le m \le M, 1 \le t \le T \end{aligned} \tag{4.8}$$

M-step: In this step, we estimate the parameters of the new GMM $\lambda^{(i+1)} = \left\{ p_m^{(i+1)}, \boldsymbol{\mu}_m^{(i+1)}, \boldsymbol{\Sigma}_m^{(i+1)} \right\}$, $m = 1, 2, \ldots, M$ using the following formulae:

Mixture Weights:

$$p_m^{(i+1)} = \frac{1}{T}\sum_{t=1}^{T}\gamma_{t,m}, \quad 1 \le m \le M \tag{4.9}$$

Means:

$$\boldsymbol{\mu}_m^{(i+1)} = \frac{\sum_{t=1}^{T}\gamma_{t,m}\mathbf{x}_t}{\sum_{t=1}^{T}\gamma_{t,m}}, \quad 1 \le m \le M \tag{4.10}$$

Covariances:

$$\boldsymbol{\Sigma}_m^{(i+1)} = \frac{\sum_{t=1}^{T}\gamma_{t,m}\left(\mathbf{x}_t - \boldsymbol{\mu}_m^{(i+1)}\right)\left(\mathbf{x}_t - \boldsymbol{\mu}_m^{(i+1)}\right)^T}{\sum_{t=1}^{T}\gamma_{t,m}}, \quad 1 \le m \le M \tag{4.11}$$

In the case of diagonal covariances, defining $\mathbf{x}^2 = diag(\mathbf{x}\mathbf{x}^T)$ the above equation becomes:

$$\boldsymbol{\sigma}_m^{2(i+1)} = \frac{\sum_{t=1}^{T}\gamma_{t,m}\mathbf{x}_t^{\,2}}{\sum_{t=1}^{T}\gamma_{t,m}} - \boldsymbol{\mu}_m^{2(i)}, \quad 1 \le m \le M \tag{4.12}$$

The new GMM, $\lambda^{(i+1)}$, becomes the initial model for the next iteration, and this process is repeated until some convergence threshold is reached. At most 200 iterations are performed.

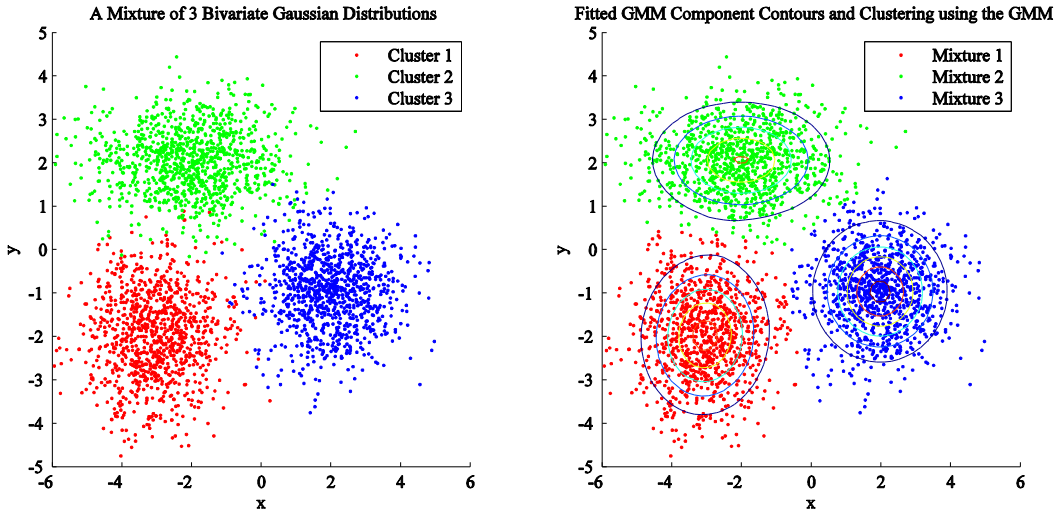

Fig. 4.2: An example of fitting a GMM to two dimensional data.

In order to visualize how Gaussian Mixture modeling works, a simple test was conducted using randomly generated two-dimensional data. The data was created using three bi-variate Gaussian distributions, as shown in Fig. 4.2(a). A GMM with *M=3* mixture components was built using the data. The contour plot of the probability density function of the GMM is shown in Fig. 4.2(b). Each data point is allocated to the mixture component with which its membership weight is highest.

From Fig. 4.2, it is clear that the Gaussian Mixture Model is able to closely approximate the actual distribution of the data. Each mixture component of the GMM seems to be modeling the distribution of one of the clusters in the data.

4.2.2.3 Initialization using K-Means Clustering

The EM algorithm described above must be initialized with some starting model in the first iteration, given by, $\lambda^{(1)} = \left\{ p_m^{(1)}, \boldsymbol{\mu}_m^{(1)}, \boldsymbol{\Sigma}_m^{(1)} \right\}$, $m = 1, 2, \ldots, M$. Irrespective of the choice of starting model, the EM algorithm will converge to some local maximum likelihood model. Since the likelihood equation has several local maxima, different initializations would cause the algorithm to converge to different maxima. However, the difference between the final models is insignificant in terms of speaker identification performance [7]. The effect of various initialization techniques such as random initialization, k-means clustering, etc. have been investigated previously. Although the use of clustering techniques for initialization does not impact performance, it provides a better starting guess compared to a random initialization. This leads to convergence of the EM algorithm in fewer iterations. In our speaker identification system, we have used k-means clustering to determine the starting model parameters.

Given a set of vectors, $X = \left\{ \mathbf{x}_t \in \mathbb{R}^D : 1 \le t \le T \right\}$, k-means clustering can be used to partition the set into *M* exclusive clusters, $X_1, X_2, ..., X_M$ with cluster centroids $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_M$ such that the within cluster sum of squares, *D* is minimized, i.e.

$$D = \sum_{m=1}^{M} \sum_{\mathbf{x}_t \in X_m} \left\| \mathbf{x}_t - \boldsymbol{\alpha}_m \right\|^2 \tag{4.13}$$

The standard algorithm for k-means clustering is an iterative technique known as Lloyd's algorithm. The algorithm is initialized by randomly choosing *M* vectors from the set *X* as cluster centroids. In the i^{th} iteration, the algorithm begins with an initial estimate of the cluster centroids, $\boldsymbol{\alpha}_1^{(i)}, \boldsymbol{\alpha}_2^{(i)}, ..., \boldsymbol{\alpha}_M^{(i)}$ and computes an improved estimate of the cluster centroids, given by $\boldsymbol{\alpha}_1^{(i+1)}, \boldsymbol{\alpha}_2^{(i+1)}, ..., \boldsymbol{\alpha}_M^{(i+1)}$. These new cluster centroids become the initial estimates in the next iteration.

Each iteration consists of an assignment step, followed by an update step, as described below.

Assignment: Each vector is assigned to the cluster that has the closest centroid, i.e.

$$X_m^{(i)} = \left\{ \mathbf{x}_t : \left\| \mathbf{x}_t - \boldsymbol{\alpha}_m^{(i)} \right\|^2 \le \left\| \mathbf{x}_t - \boldsymbol{\alpha}_n^{(i)} \right\|^2 \ \forall n, 1 \le n \le M, 1 \le t \le T \right\}, \quad 1 \le m \le M \tag{4.14}$$

Each vector is assigned to exactly one cluster, even if it could be assigned to more than one of them.

Update: In this step, the cluster centroids are re-estimated to be the centroids of the new clusters.

$$\boldsymbol{\alpha}_m^{(i+1)} = \frac{1}{\left| X_m^{(i)} \right|} \sum_{\mathbf{x}_t \in X_m^{(i)}} \mathbf{x}_t, \quad 1 \le m \le M \tag{4.15}$$

where $\left| X_m^{(i)} \right|$ is the cardinality of the set $X_m^{(i)}$. The k-means algorithm is said to have converged when the cluster assignments no longer change. The iterations are continued until convergence, or until the maximum of 200 iterations is reached.

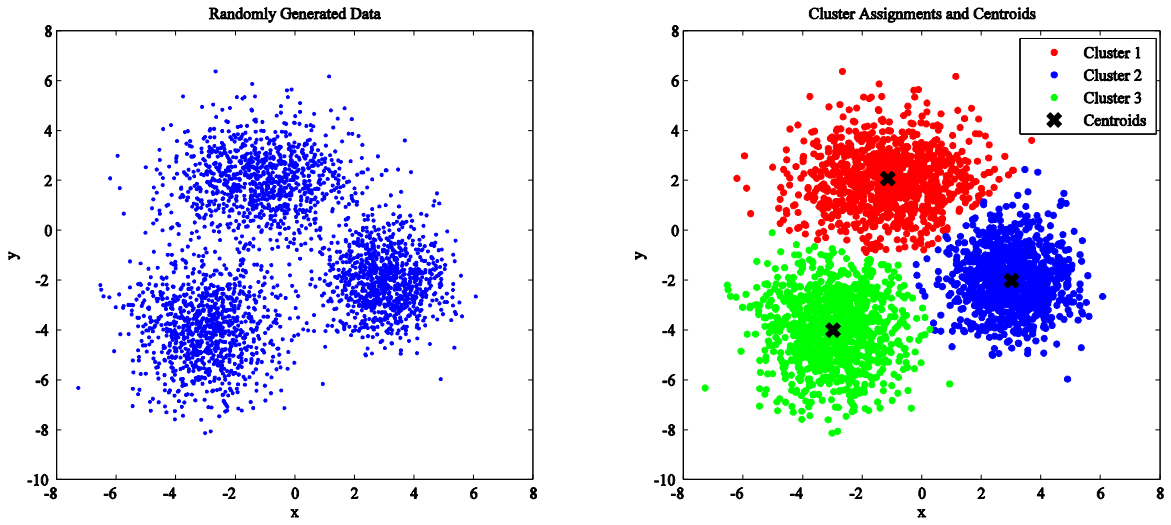

Fig. 4.3: An example of clustering using the k-means algorithm.

In order to visualize how k-means clustering works, a simple test was conducted using randomly generated two-dimensional data from three different bi-variate Gaussian distributions. The k-means clustering algorithm was used to partition the data into three clusters. The cluster assignments and centroid locations after k-means clustering are shown in Fig. 4.3.

Let us denote the final clusters and cluster centroids after k-means clustering by $X_1, X_2, ..., X_M$ and $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_M$ respectively.

Using these results, we can estimate the parameters of the starting GMM $\lambda^{(1)} = \left\{ p_m^{(1)}, \boldsymbol{\mu}_m^{(1)}, \boldsymbol{\Sigma}_m^{(1)} \right\}, m = 1, 2, \ldots, M$ as follows:

$$p_m^{(1)} = \frac{|X_m|}{\sum_{m=1}^{M} |X_m|}, \quad 1 \le m \le M \tag{4.16}$$

$$\boldsymbol{\mu}_m^{(1)} = \boldsymbol{\alpha}_m, \quad 1 \le m \le M \tag{4.17}$$

Then, the covariance matrix of each individual component can be initialized as:

$$\boldsymbol{\Sigma}_m^{(1)} = \frac{1}{|X_m| - 1} \sum_{\mathbf{x}_t \in X_m} (\mathbf{x}_t - \boldsymbol{\alpha}_m)(\mathbf{x}_t - \boldsymbol{\alpha}_m)^T, \quad 1 \le m \le M \tag{4.18}$$

For reasons explained earlier, we restrict ourselves to diagonal covariance matrices in our speaker identification system. So, we create a diagonal covariance matrix by retaining only the variances on the main diagonal of the matrix in (4.18) and use it for initialization.

Next, the effect of using k-means based initialization instead of random initialization is investigated. A GMM with *M=3* components is fitted to randomly generated two-dimensional data from three bi-variate Gaussian distributions.

From Fig. 4.4, we can see that when the EM algorithm uses random initialization, it takes 20 iterations to converge. On the other hand, the EM algorithm converges in 7 iterations using k-means initialization. However, the likelihood of the final GMM is nearly the same for both of these initialization techniques.

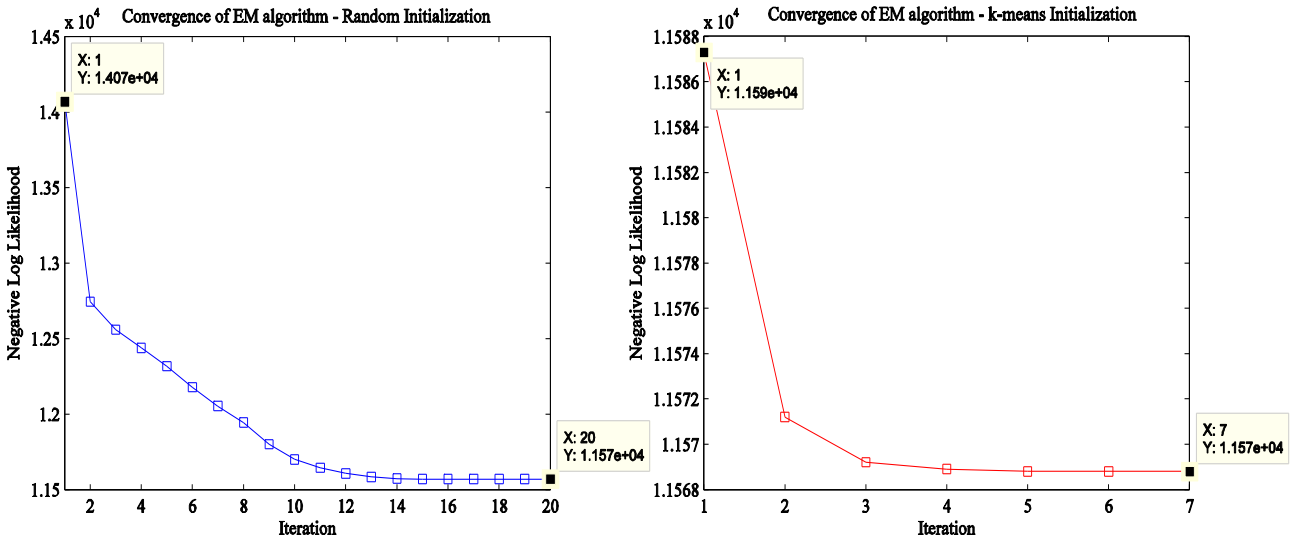

Fig. 4.4: Effect of initialization method on convergence of the EM algorithm.

4.2.3 Speaker Identification

Upon completion of the training phase, we would have built *S* Gaussian Mixture Models, $\lambda_1, \lambda_2, ..., \lambda_S$, one for each of the *S* enrolled speakers. The identification or testing phase is illustrated in the following figure.

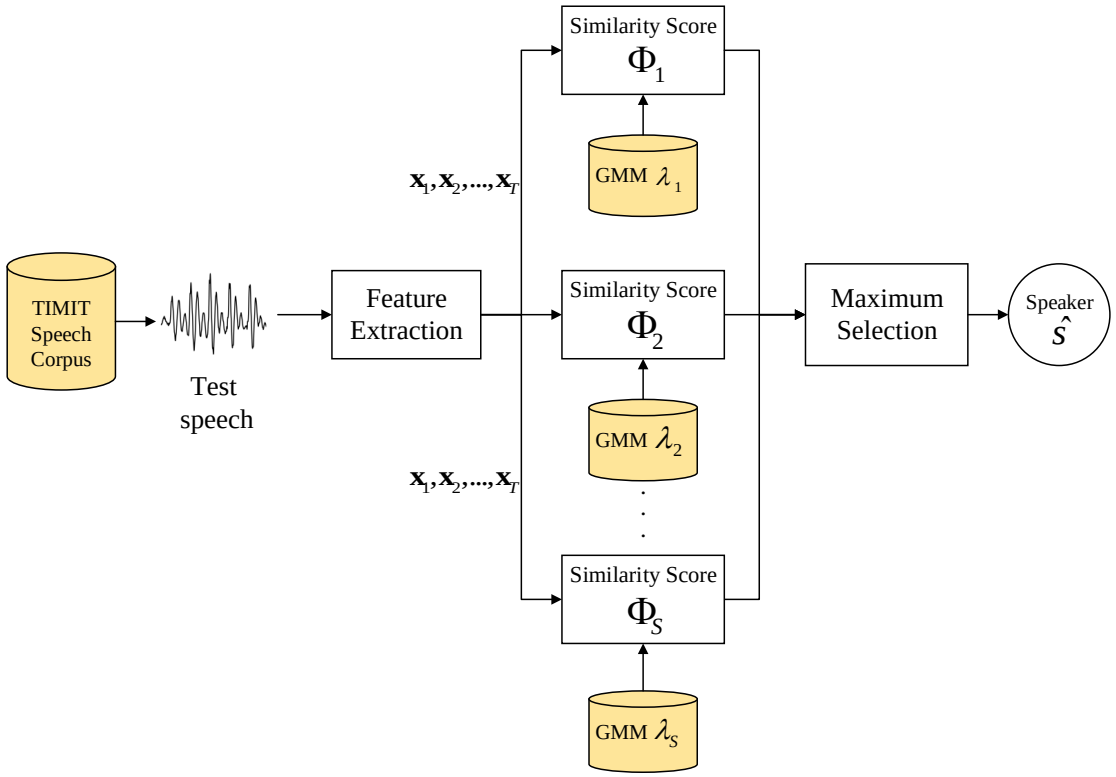

Fig. 4.5: Speaker identification process.

During speaker identification, feature vectors are extracted from the unknown speaker's test utterance. Let us assume that a set of *T* feature vectors of dimension *D* was extracted from the unknown speaker's test utterance, given by:

$$X = \left\{ \mathbf{x}_t \in \mathbb{R}^D : 1 \le t \le T \right\} \tag{4.19}$$

The objective of speaker identification is to find that enrolled speaker whose speaker model has the maximum a posteriori probability, given a set of test features from an unknown speaker. This can be written as:

$$\hat{s} = \arg\max_{1 \le s \le S} p(\lambda_s \mid X) \tag{4.20}$$

where $\hat{s}$ is the identity of the unknown speaker determined by the speaker identification system.

Using Bayes' rule, the above expression becomes:

$$\hat{s} = \arg\max_{1 \le s \le S} \frac{p(X \mid \lambda_s)\, p(\lambda_s)}{p(X)} \tag{4.21}$$

Assuming equally likely speakers, we can say that $p(\lambda_s) = \frac{1}{S}$. Also, $p(X)$ is the same for all speaker models, and so we get:

$$\hat{s} = \arg\max_{1 \le s \le S} p(X \mid \lambda_s) \tag{4.22}$$

Using the fact that $p(X \mid \lambda_s) = \prod_{t=1}^{T} p(\mathbf{x}_t \mid \lambda_s)$ in (4.22), we get:

$$\hat{s} = \arg\max_{1 \le s \le S} \prod_{t=1}^{T} p(\mathbf{x}_t \mid \lambda_s) \tag{4.23}$$

If we use log-probabilities instead, the product term above would become a sum of the log-probabilities. Since the logarithm function is monotonically increasing, we can write:

$$\hat{s} = \arg\max_{1 \le s \le S} \sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \lambda_s) \tag{4.24}$$

Now, we can define the *similarity score* of the unknown speaker with the speaker model λ_s as:

$$\Phi_s = \sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \lambda_s) \tag{4.25}$$

Substituting (4.25) in (4.24) we get:

$$\hat{s} = \arg\max_{1 \le s \le S} \Phi_s \tag{4.26}$$

Thus, the speaker whose speaker model gives the maximum similarity score with the test utterance is determined to be the unknown speaker.

We can also define the *frame score* of the t^{th} frame with the speaker model λ_s to be:

$$\phi_{t,s} = \log p(\mathbf{x}_t \mid \lambda_s) \tag{4.27}$$

Then, the similarity score can be written as a sum of the T frame scores:

$$\Phi_s = \sum_{t=1}^{T} \phi_{t,s} \tag{4.28}$$

Recently, a *weighted scoring* mechanism [79] has been proposed, in which the t^{th} frame, or equivalently feature, in the test utterance is given a weight w_t, $0 \le w_t \le 1$. The similarity score is then computed as:

$$\Phi_s = \frac{\sum_{t=1}^{T} w_t \phi_{t,s}}{\sum_{t=1}^{T} w_t} \tag{4.29}$$

4.2.4 Performance Evaluation

The performance of an SI system is evaluated in the following manner. In each test utterance, the speaker identity determined by the system is compared to the true speaker identity. The number of tests in which a speaker was correctly identified is counted. The final performance measure is then the percentage of tests in which the system identified the speaker correctly. This measure is known as the Identification Accuracy:

$$IA(\%) = \frac{\#\ of\ \ correctly\ identified\ tests}{total\ \#\ of\ \ tests} \tag{4.30}$$

4.3 SPEAKER VERIFICATION SYSTEM

In this section, the enrollment and testing (or verification) phases of our speaker verification (SV) system are described. As explained in Chapter 1, speaker verification is a two-class decision task. Given a set of feature vectors $X = \{\mathbf{x}_t \in \mathbb{R}^D : 1 \le t \le T\}$ coming from a test utterance, speaker verification can be thought of as a basic hypothesis test between:

H_0: X is from the hypothesized speaker s, and H_1: X is not from the hypothesized speaker s

The null hypothesis H_0 is represented by the speaker model of the hypothesized speaker, denoted by λ_s. On the other hand, the alternate hypothesis H_1 must represent the entire space of possible alternatives to the hypothesized speaker, and is denoted by λ_{alt}. The optimum test to decide between the hypotheses is a likelihood ratio test, given by:

$$\frac{p(X \mid \lambda_s)}{p(X \mid \lambda_{alt})} \begin{cases} \ge \theta & accept\ H_0 \\ < \theta & reject\ H_0 \end{cases} \tag{4.31}$$

Taking the logarithm of this statistic, we get:

$$\Lambda(X) = \log p(X \mid \lambda_s) - \log p(X \mid \lambda_{alt}) \tag{4.32}$$

The alternative hypothesis model λ_{alt} is created by pooling speech from a large population of background speakers and training a single Gaussian Mixture Model. Such a model is known as a Universal Background Model (UBM), and is denoted by λ_{ubm}.

4.3.1 Universal Background Model

In our SV system, the Universal Background Model (UBM) λ_{ubm} is a GMM created by pooling speech from all the 462 speakers in the TRAIN directory of the TIMIT corpus. All the 10 utterances from each of the 462 speakers are utilized for training the UBM. Using features extracted from the pooled speech, a GMM with 256 mixture components and diagonal covariance matrices is trained using the MSR Identity Toolbox [97].

A binary splitting procedure is used to train the UBM, starting from a single component to 256 components. All training vectors are initially placed in a single cluster and the mean is calculated. In each split, the mean of each cluster is perturbed by a small value ε along the dimension associated with the maximum variance, giving rise to two new clusters:

$$\begin{aligned} \boldsymbol{\mu}_i^+ &= \boldsymbol{\mu}_i(1+\varepsilon) \\ \boldsymbol{\mu}_i^- &= \boldsymbol{\mu}_i(1-\varepsilon) \end{aligned} \tag{4.33}$$

After each split, the model is re-estimated several times using the EM algorithm. The number of EM iterations at each split is gradually increased from 1 to 10 for the 256 component GMM.

4.3.2 Training and Test Sets

Similar to the SI system, all SV experiments are conducted using the 168 speakers in the TEST directory of the TIMIT corpus. The speech material of each speaker in the TEST directory is divided into mutually exclusive training and test sets. The training set consists of the 5 SX sentences and the 3 SI sentences. The test set consists of the two remaining SA sentences.

4.3.3 Speaker Enrollment Using Adapted GMM

In our SV system, a speaker model is created for each speaker by adapting the parameters of the UBM, using speech from the speaker's training set. This is also known as Bayesian adaptation or maximum a posteriori (MAP) adaptation. This provides a tighter link between each speaker's model and UBM, thereby improving performance and allowing for faster scoring [98]. Given a set of feature vectors obtained from the training speech, $X = \left\{ \mathbf{x}_t \in \mathbb{R}^D : 1 \le t \le T \right\}$, we compute the a posteriori probability of each feature vector belonging to each of the *M* Gaussian components in the UBM.

$$\begin{aligned}\gamma_{t,m} &= p\left(m \mid \mathbf{x}_t, \lambda_{ubm}\right) \\ &= \frac{p_m g_m\left(\mathbf{x}_t \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right)}{\sum_{k=1}^{M} p_k g_k\left(\mathbf{x}_t \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right)}, \quad 1 \le m \le M, 1 \le t \le T\end{aligned} \tag{4.34}$$

Next, we compute some sufficient statistics, namely the count, first, and second moments required to compute the mixture weights, mean, and variance.

$$n_m = \sum_{t=1}^{T} \gamma_{t,m}, \quad 1 \le m \le M \tag{4.35}$$

$$E_m(\mathbf{x}) = \frac{\sum_{t=1}^{T} \gamma_{t,m} \mathbf{x}_t}{\sum_{t=1}^{T} \gamma_{t,m}}, \quad 1 \le m \le M \tag{4.36}$$

$$E_m(\mathbf{x}^2) = \frac{\sum_{t=1}^{T} \gamma_{t,m} \mathbf{x}_t^{\,2}}{\sum_{t=1}^{T} \gamma_{t,m}}, \quad 1 \le m \le M \tag{4.37}$$

These new sufficient statistics from the training data are used to update the old UBM sufficient statistics, creating the adapted GMM mixture weights, means and variances:

$$\hat{p}_m = \left[\frac{\alpha_m^p n_m}{T} + (1 - \alpha_m^p) p_m \right] \gamma, \quad 1 \le m \le M \tag{4.38}$$

$$\hat{\boldsymbol{\mu}}_m = \alpha_m^{\mu} E_m(\mathbf{x}) + (1 - \alpha_m^{\mu}) \boldsymbol{\mu}_m, \quad 1 \le m \le M \tag{4.39}$$

$$\hat{\boldsymbol{\sigma}}_m^2 = \alpha_m^{\sigma} E_m(\mathbf{x}^2) + (1 - \alpha_m^{\sigma})(\boldsymbol{\sigma}_m^2 + \boldsymbol{\mu}_m^2) - \hat{\boldsymbol{\mu}}_m^2, \quad 1 \le m \le M \tag{4.40}$$

The scale factor, γ is computed over all adapted mixture weights to ensure they sum to unity, i.e. $\sum_{m=1}^{M} \hat{p}_m = 1$. Here $\{\alpha_m^p, \alpha_m^{\mu}, \alpha_m^{\sigma}\}$ are the adaptation coefficients for the weights, means and variances. Each adaptation coefficient is defined as:

$$\alpha_m^{\rho} = \frac{n_m}{n_m + r^{\rho}} \tag{4.41}$$

where $\rho \in \{p, \mu, \sigma\}$ and r^{ρ} is a fixed relevance factor for parameter ρ.

Previous research has shown that adapting only the means results in the best verification performance [98]. Thus, in our SV system, only the means are adapted, using a relevance factor r=16, as suggested in [98].

4.3.4 Speaker Verification

Upon completion of the training phase, we would have built S adapted Gaussian Mixture Models, $\lambda_1, \lambda_2, ..., \lambda_S$, one for each of the S enrolled speakers and a Universal Background Model (UBM), λ_{ubm}. The verification or testing phase is illustrated in Fig. 4.6. Let us assume that a set of T feature vectors of dimension D is extracted from the unknown speaker's test utterance, given by $X = \{\mathbf{x}_t \in \mathbb{R}^D : 1 \le t \le T\}$.

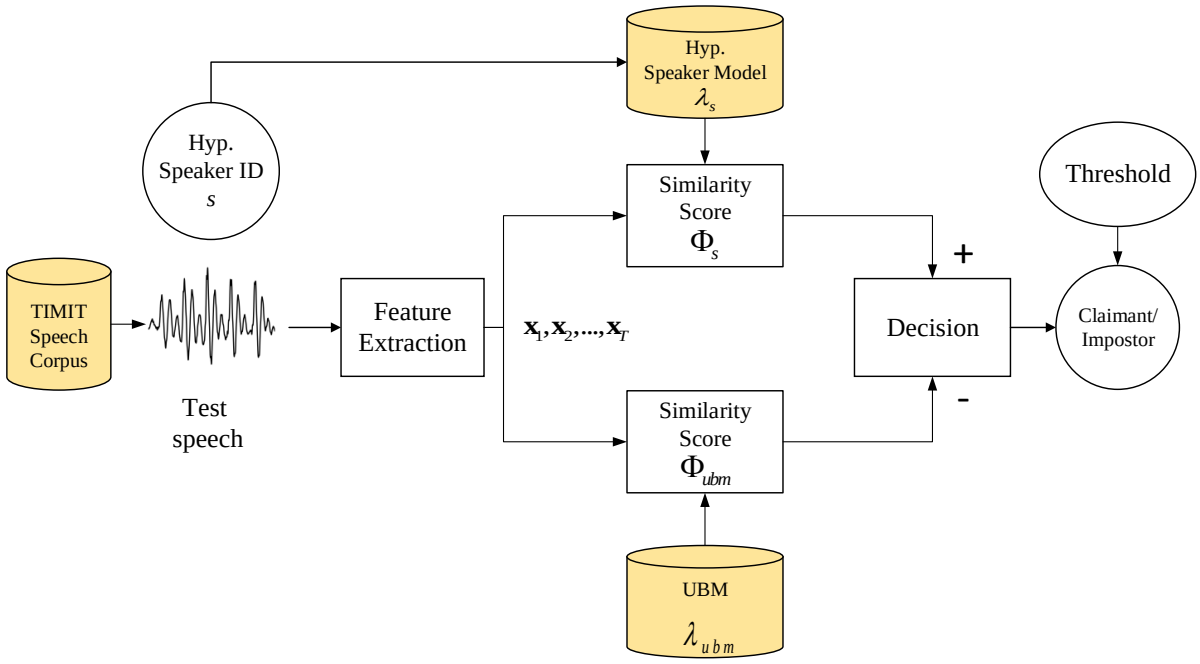

Fig. 4.6: Speaker Verification phase.

In speaker verification, the unknown speaker makes an identity claim. Let us say that he claims to be speaker *s*. Then, the speaker verification task is a simple hypothesis test between:

$H_0: X$ is from the hypothesized speaker s, and $H_1: X$ is not from the hypothesized speaker s.

The decision is made using the log likelihood ratio given by:

$$\Lambda(X) = \Phi_s - \Phi_{ubm} \tag{4.42}$$

where Φ_s and Φ_{ubm} are the similarity scores of the feature vectors with the hypothesized speaker model and the UBM respectively, given by:

$$\Phi_s = \sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \lambda_s) \tag{4.43}$$

$$\Phi_{ubm} = \sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \lambda_{ubm}) \tag{4.44}$$

If the log-likelihood ratio exceeds the selected decision threshold θ, we accept the null hypothesis H_0. Else, the null hypothesis is rejected.

4.3.5 Performance Evaluation

The performance of an SV system is evaluated in the following manner. A series of tests, known as target trials and impostor trials are conducted. All test utterances in which the hypothesized speaker is the true speaker are called target trials. All test utterances in which the hypothesized speaker is not the true speaker are called impostor trials. During verification, the scores from the target trials are pooled into a set of target scores. The scores from the impostor trials are pooled into a set of impostor scores. A single, speaker independent threshold θ is swept over the two sets of scores and the probability of miss and probability of false alarm are computed for each threshold.

The miss probability corresponds to the probability of rejecting the null hypothesis when the true speaker is actually the hypothesized speaker. In other words, it is the probability of not detecting the target speaker when present [8]. It is given by:

$$E_{miss} = n_{miss} / n_t \tag{4.45}$$

where n_t and n_{miss} are the number of target trials and the number of those where the target speaker was not detected.

The false alarm probability is the probability of accepting the null hypothesis when the true speaker is not the hypothesized speaker. In other words, it is the probability of falsely detecting the target speaker when not present. It is calculated as:

$$E_{fa} = n_{fa} / n_i \tag{4.46}$$

where n_i and n_{fa} are the number of impostor trials and the number of those where the target speaker was falsely detected, respectively.

The miss and false alarm probabilities depend on the decision threshold. By varying the value of θ over a range, we can plot the miss probability as a function of false alarm probability to show system performance. This is known as a Receiver Operator Characteristic (ROC) curve [99].

The trade-off between miss and false alarm is most commonly depicted using the detection error trade-off (DET) curve as opposed to an ROC curve. In the DET curve, we plot the miss and false alarm probabilities according to their corresponding Gaussian deviates, rather than the probabilities themselves. This results in a non-linear probability scale, but the plots are visually more intuitive. If the distributions of error probabilities are Gaussian, then the resulting trade-off curves are straight lines [8]. The distance between curves depicts differences in performance.

The main performance measure of an SV system is the *Equal Error Rate* (EER) percentage. This is essentially the operating point at which the system achieves a balanced performance, i.e. the probability value at which the miss and false alarm probabilities are equal.

5 VOWEL ONSET AND END POINT DETECTION

In Chapter 2, we presented evidence suggesting that transient zones of speech around phoneme boundaries encapsulate speaker-specific characteristics. The objective of this thesis is to investigate the speaker-discriminative power of these regions under noisy conditions. In particular, we focus on extracting features from transitions into and out of vowels, i.e. at consonant-vowel boundaries. We select only transitions into/out of vowels because vowels are easier to locate due to their higher energy, especially in noisy conditions.

In order to isolate the transitions into and out of vowels, we propose to first locate the vowel regions in speech. In a typical consonant-vowel-consonant (CVC) syllable in speech, the end of the consonant part and the beginning of the vowel part is called the Vowel Onset Point (VOP). Similarly, the Vowel End Point (VEP) marks the end of the vowel region. In the region of 20 milliseconds before and after a VOP or VEP, the dynamics of the speech signal vary rapidly. Thus, once the VOPs and VEPs have been determined, we can locate transition frames by examining a small window around them.

The VOP and VEP detection algorithm used in our research has been adapted from the methods proposed by Prasanna et al. [82, 83] and Pradhan et al. [76]. This algorithm uses complementary information from the Hilbert Envelope of the LP residual (HE-LP), the Zero Frequency Filtered Signal (ZFFS) and the Spectral Peaks Energy (SPE) to identify vowel onset and offset points.

5.1 HILBERT ENVELOPE OF THE LP RESIDUAL

In linear prediction (LP) analysis, it is observed that the excitation signal, also known as the LP residual, is maximal around the glottal closure instants (GCIs) [76]. Since major glottal activity takes place during the vowel regions, the LP residual is also expected to be greater in these regions. Thus, the LP residual could be used to identify vowel regions in a speech signal.

To speed up the processing, the speech signal is first downsampled to 8 kHz. Then, the signal is passed through a pre-emphasis filter with $\alpha = 0.97$. Next, the speech signal is blocked into 20 msec long frames, with 10 msec frame shift, and each frame is multiplied with a Hamming window. In each frame, linear prediction analysis of order 10 is used to find the inverse vocal tract filter $A_p(z)$, as explained in Chapter 3.

The LP residual, or the excitation signal $e[n]$, in each frame is computed by passing the windowed speech frame $x[n]$ through the inverse filter $A_p(z)$. Once the frame LP residuals have been obtained, the LP residual of the entire speech signal, $r[n]$ can be constructed by time-aligning the frame LP residuals and overlap-adding. Due to the bipolar nature of the LP residual, the time-

varying changes are not clearly visible. In order to enhance these changes, the Hilbert Envelope of the LP Residual (HE-LP) is calculated:

$$h[n] = \sqrt{r^2[n] + r_h^2[n]} \tag{5.1}$$

where $r_h[n]$ is the Hilbert Transform of $r[n]$. Since we want to preserve only variations at a pitch period level, $h[n]$ is smoothed by taking the maximum value of $h[n]$ in every 10 msec block with one sample shift [75]. The HE-LP signal after smoothing is denoted by $h_s[n]$. As the vowel regions correspond to large values of the HE-LP, we can identify vowel onset and offset points by looking for gross amplitude changes in $h_s[n]$. This is accomplished by convolving $h_s[n]$ with a first-order Gaussian differentiator (FOGD), shown in Fig. 5.1. An FOGD of length L and standard deviation σ is written as:

$$g'[n] = g[n] - g[n-1] \tag{5.2}$$

where $g[n]$ is a Gaussian window of length $L+1$ and standard deviation σ given by:

$$g[n] = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(n-\frac{L}{2}\right)^2}{2\sigma^2}}, \quad 0 \le n \le L \tag{5.3}$$

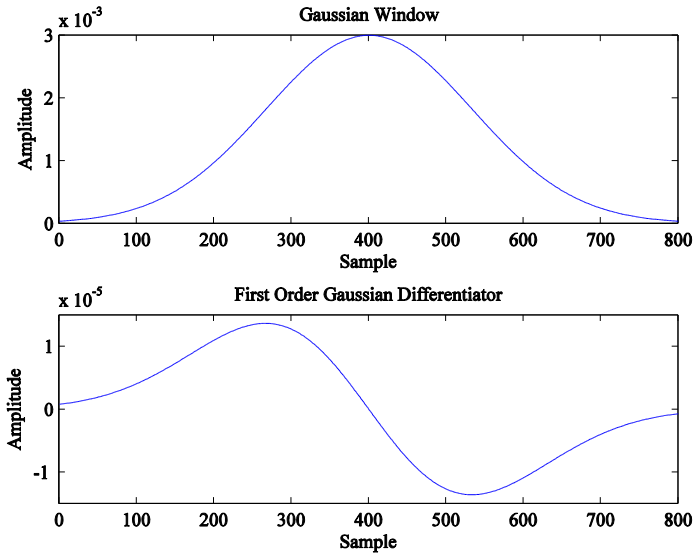

Fig. 5.1: A Gaussian window and the corresponding FOGD.

To obtain the VOP evidence $v_o[n]$, $h_s[n]$ is convolved with a FOGD $g_o'[n]$, with a length of 100 msec ($L = 800$) and a standard deviation of one-sixth of the window length ($\sigma = 134$), from left to right. The parameters of the FOGD are chosen assuming that VOPs occur as sharp level changes at intervals of around 100 msec.

$$v_o[n] = h_s[n] * g'_o[n] \tag{5.4}$$

The signal characteristics at a vowel offset are significantly different from those at the vowel onset. At the onset, there is a sudden increase in signal strength, while the signal strength decreases slowly at the offset. Due to this, we convolve $h_s[n]$ from right to left with an FOGD $g'_e[n]$ which is double in length and standard deviation compared to $g'_o[n]$ as shown below.

$$v_{er}[n] = h_s[-n] * g'_e[n] \tag{5.5}$$

Then, the VEP evidence $v_e[n]$ is:

$$v_e[n] = v_{er}[-n] \tag{5.6}$$

A speech signal and its LP residual are shown in Fig. 5.2(a)-(b). The Hilbert envelope of the LP residual (HE-LP) is shown in Fig. 5.2(c). The HE-LP signal after smoothing is shown in Fig. 5.2(d). By observing Fig. 5.2(e)-(f), we see that sharp peaks are obtained in the VOP/VEP evidence when there is a sharp amplitude increase/decrease in the smoothed HE-LP respectively.

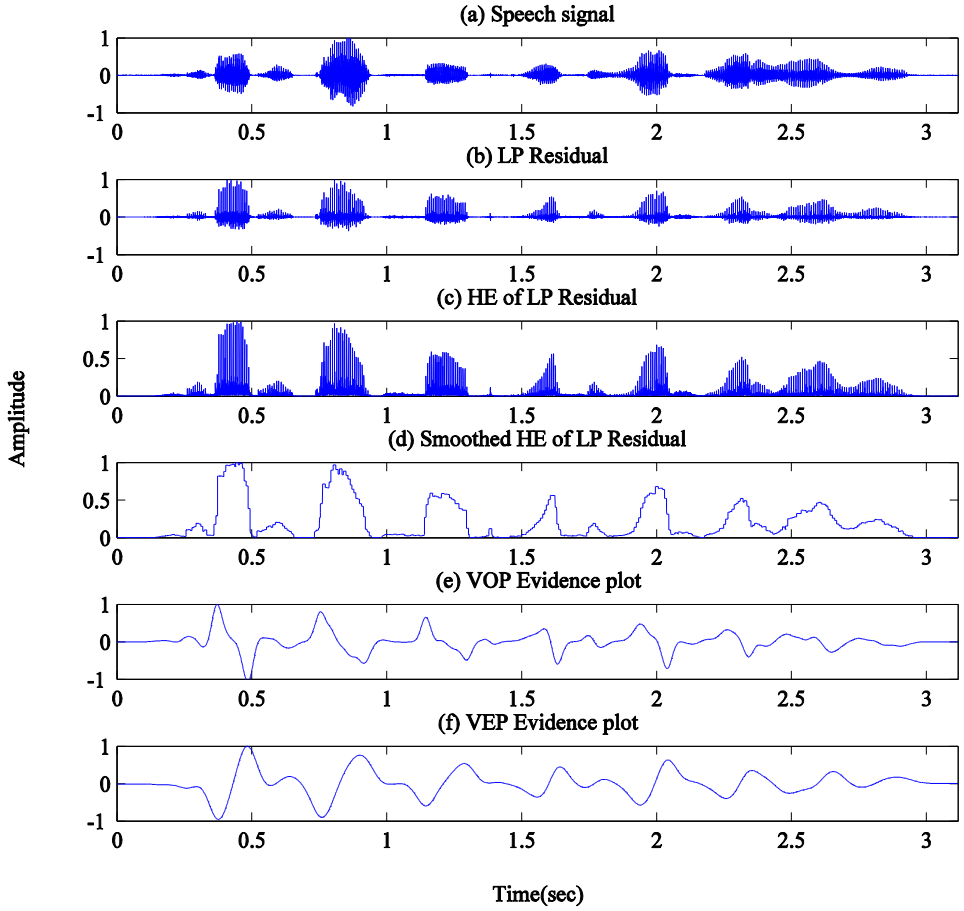

Fig. 5.2: VOP/VEP Evidence from the Hilbert Envelope of the LP Residual.

5.2 ZERO FREQUENCY FILTERED SIGNAL

Vowel regions in a speech signal are produced by a periodic vibration of the vocal folds. As explained earlier, during vowel regions, the excitation signal is a series of periodic glottal pulses, much like an impulse train. The time-domain representation of an impulse has an equivalent frequency domain representation of impulses uniformly located at all the frequencies including zero frequency, separated by the fundamental frequency.

The effect of the impulse-like excitation is clearly visible upon filtering the signal with a narrowband filter, i.e. a near ideal resonator at zero frequency [100]. This preserves only the signal energy around the impulse present at zero frequency and removes all high frequency information, mainly due to the vocal tract resonances. The signal obtained by passing the speech through a zero frequency resonator, is known as the Zero Frequency Filtered Signal (ZFFS). Before extracting the ZFFS, the speech signal is first downsampled to 8 kHz, and then differentiated twice to remove any slowly varying components.

$$s''[n] = s[n+1] - 2s[n] + s[n-1] \tag{5.7}$$

An ideal zero-frequency digital resonator is an IIR filter with a pair of poles located on the unit circle. In order to provide a sharper cut-off, a cascade of two such resonators is used [101].

$$y[n] = -\sum_{k=1}^{4} a_k y[n-k] + s''[n] \tag{5.8}$$

where $a_1 = -4, a_2 = 6, a_3 = -4, a_4 = 1$. Next, the trend in $y[n]$ is removed by subtracting the local mean computed over a 10 msec window (corresponding to around two pitch periods), at each sample. The trend removed signal, shown in Fig. 5.3(b), is termed the Zero Frequency Filtered Signal (ZFFS), and is denoted by $z[n]$.

$$\bar{y}[n] = \frac{1}{2N+1} \sum_{m=-N}^{N} y[n+m] \tag{5.9}$$

$$z[n] = y[n] - \bar{y}[n] \tag{5.10}$$

As illustrated in Fig. 5.3(c), the time-varying changes in the ZFFS are enhanced further by computing the absolute value of its second-order difference, given by:

$$z_d[n] = \left| z[n+1] - 2z[n] + z[n-1] \right| \tag{5.11}$$

Since we want to preserve only variations at a pitch period level, $z_d[n]$ is smoothed by taking the maximum value in every 10 msec block with one sample shift. The smoothed signal is denoted by $z_s[n]$, and is shown in Fig. 5.3(d). Next, $z_s[n]$ is convolved with the same FOGD $g_o'[n]$, from left

to right as explained in (5.4), to obtain the VOP evidence $v_o[n]$. Similarly, to obtain the VEP evidence $v_e[n]$, we convolve $z_s''[n]$ from right to left with the FOGD $g_e'[n]$, as explained in (5.6). The VOP and VEP evidences are shown in Fig. 5.3(e)-(f).

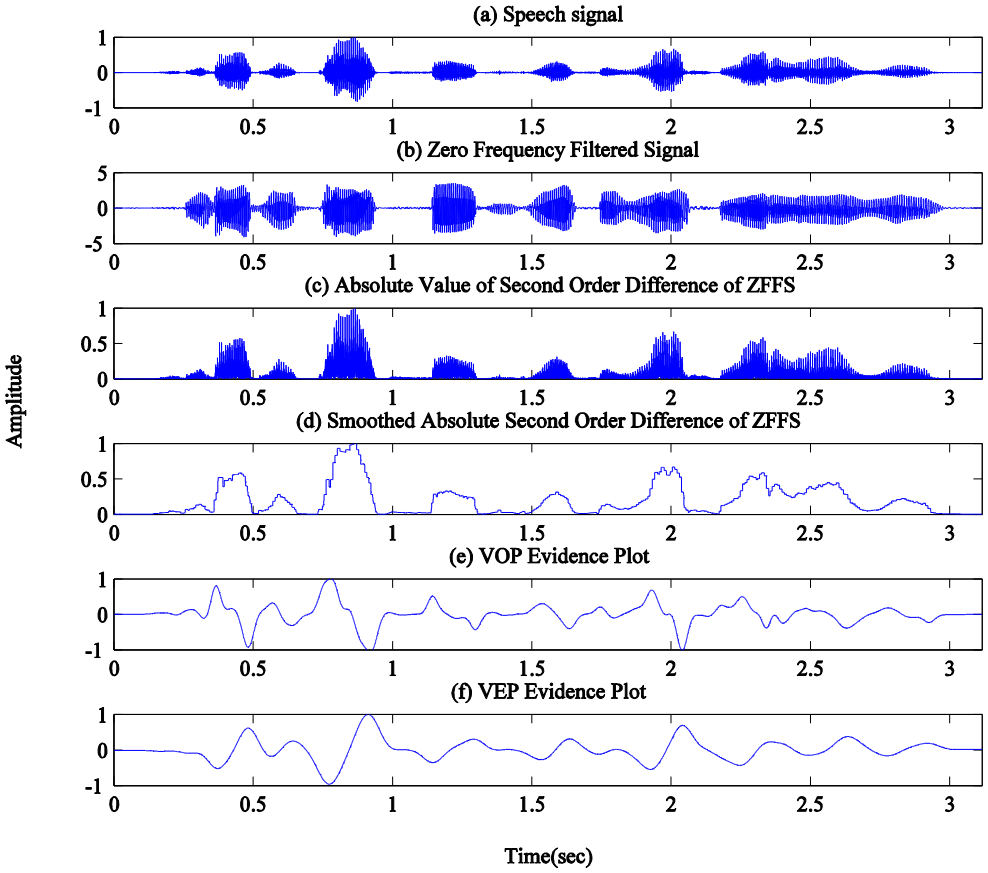

Fig. 5.3: VOP/VEP Evidence obtained using the Zero Frequency Filtered Signal.

5.3 SPECTRAL PEAKS ENERGY

Vowels are high-energy regions of speech, since they are produced by an open configuration of the vocal tract. The spectral energy of vowels primarily lies in the range 250 – 2,500 Hz and is concentrated at the formants, which correspond to the peaks of the spectrum. Thus, the energy of the spectral peaks below 2500 Hz would give us an indication of where the vowel regions lie.

The speech signal is first downsampled to 8 kHz and passed through a pre-emphasis filter with $\alpha = 0.97$. Next, the signal is passed through a low-pass FIR filter of order 100, with a passband edge frequency of 2500 Hz. The passband ripple is set to 0.01 dB and the stopband attenuation is chosen as 80 dB. The magnitude and phase response of the low-pass filter are shown in Fig. 5.4. Then, the low-pass filtered signal is divided into overlapping frames of 20 msec length, with a frame shift of 10 msec. Each frame is windowed using a Hamming window. A 256-point DFT is computed for each frame, and the ten largest peaks are selected from the first 128 points.

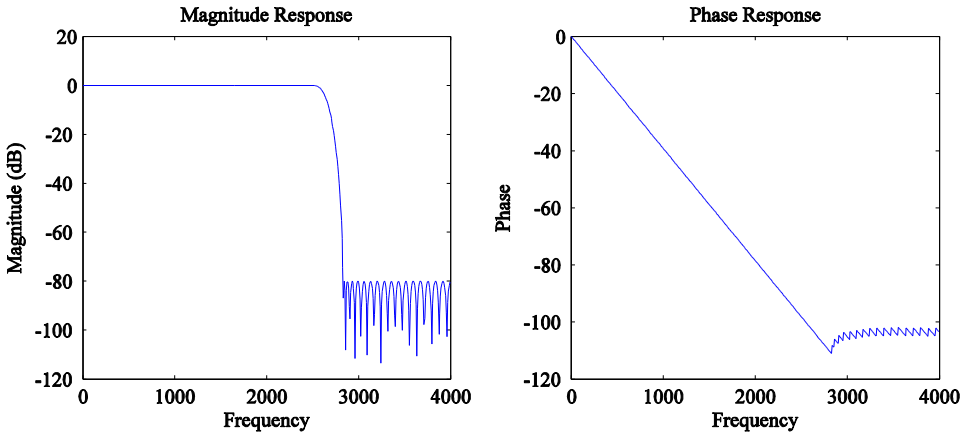

Fig. 5.4: Low pass filter with passband-edge frequency of 2500 Hz.

The sum of the energies in the ten largest DFT peaks in each frame, plotted as a function of time, is used as the representation of the spectral peaks energy. The spectral peaks energy obtained via the steps explained above is shown in Fig. 5.5(b). The spectral peaks energy is convolved with the FOGD $g_o'[n]$, from left to right, as explained in (5.4) to obtain the VOP evidence $v_o[n]$. Similarly, to obtain the VEP evidence $v_e[n]$, we convolve the spectral peaks energy from right to left with the FOGD $g_e'[n]$, as explained in (5.6). The VOP and VEP evidences are shown in Fig. 5.5(c)-(d).

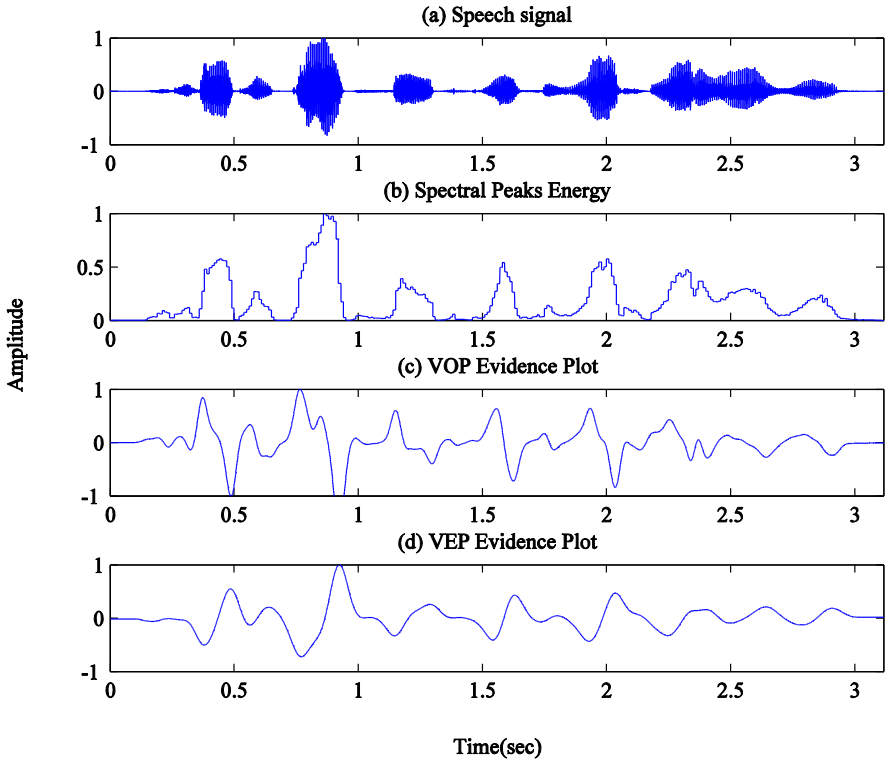

Fig. 5.5. VOP/VEP Evidence obtained using Spectral Peaks Energy.

5.4 COMBINATION OF EVIDENCES

The combined VOP evidence is obtained by adding the VOP evidences from the HE-LP and the ZFFS methods, and normalizing by the maximum value of the sum. Similarly, the combined VEP evidence is obtained by adding the VOP evidences from the HE-LP and the ZFFS methods, and normalizing by the maximum value of the sum.

The peaks in the combined VOP evidence are selected by finding the maximum value between two successive positive to negative zero crossings with some small threshold (~0.03-0.1) to eliminate the spurious ones. It has been found that the choice of this threshold is not critical. These peak locations are the hypothesized VOP candidates [76]. Since every VOP is followed by a VEP, the locations of the valley immediately after each peak are made note of.

Similarly, the peaks in the combined VEP evidence are selected by finding the maximum value between two successive positive to negative zero crossings with some small threshold (~0.03-0.08) to eliminate the spurious ones. It has been found that the choice of this threshold is not critical. The locations of these peaks are the hypothesized VEP candidates. Since every VEP is preceded by a VOP, the locations of the valley immediately before each peak are made note of.

Every vowel onset must be followed by a vowel offset, and so, we can force the detection of missing cases and reduce the spurious detections of one event using the knowledge of the other event. This is done using a simple, two-stage algorithm [76], which has been outlined below.

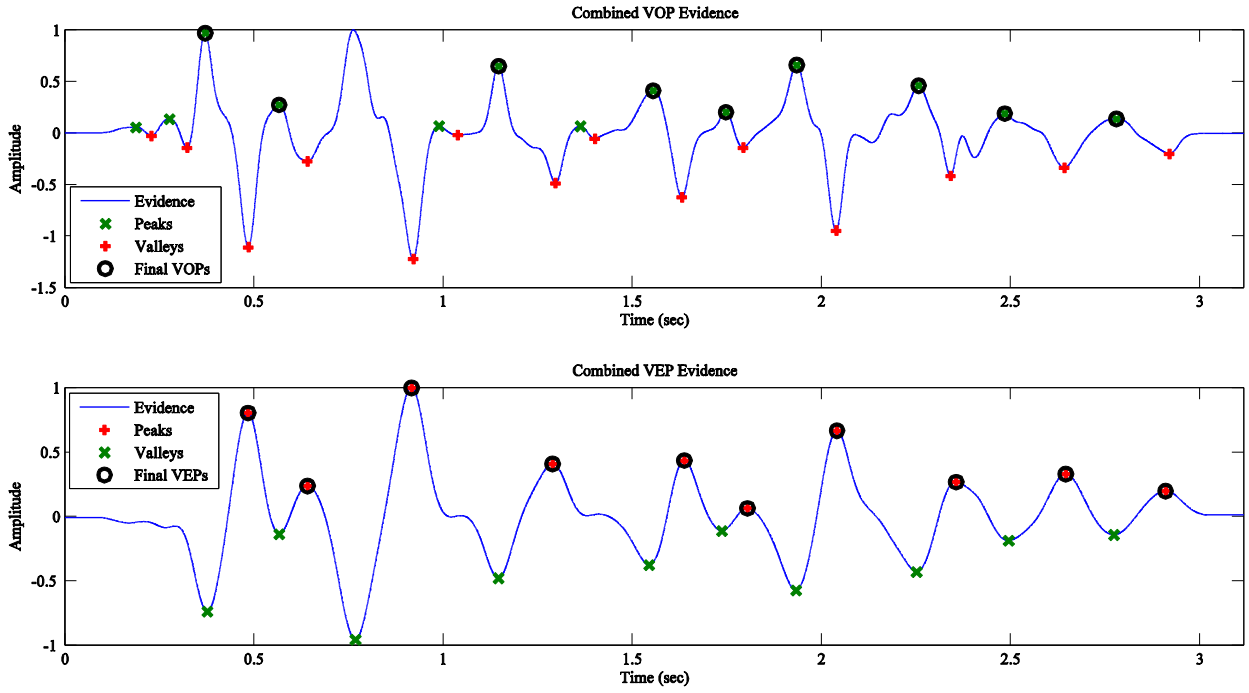

Fig. 5.6. Hypothesized VOP and VEP locations.

- **Stage 1: Forced Detection**

  In this stage, a VOP is labeled as 'strong', if the VOP evidence at this location exceeds 10% of the maximum VOP evidence. The algorithm checks to see if a VEP has been detected between two consecutive strong VOPs, from left to right. If not, the valley point in the VOP evidence bounded by the two strong VOPs is marked as a VEP.

  Similarly, a VEP is considered to be 'strong', if the VEP evidence at this location exceeds 10% of the maximum VEP evidence. The algorithm checks to see if a VOP has been detected between two consecutive strong VEPs, from right to left. If not, the valley point in the VEP evidence bounded by the two strong VEPs is marked as a VOP.

- **Stage 2: Spurious Removal**

  Next, all the VOPS hypothesized originally and the forcibly detected ones are pooled. Similarly, all the VEPS hypothesized originally and the forcibly detected ones are collected. Between two consecutive VOPs, if there is no VEP, then the VOP with the weaker evidence is removed. On the contrary, if there are two or more VEPs between two consecutive VOPs, only the VEP with the highest evidence is retained and the other VEPs are discarded. If one or more VEPs exist after the last VOP, then only the VEP with the highest evidence is retained. If there is no VEP after the last VOP, the last VOP is discarded. In addition, if there is any VEP before the first VOP, it is removed.

The VOPs and VEPs remaining at the end of the two-stage algorithm are considered to be final. The results of the VOP and VEP algorithm are shown in the following figure.

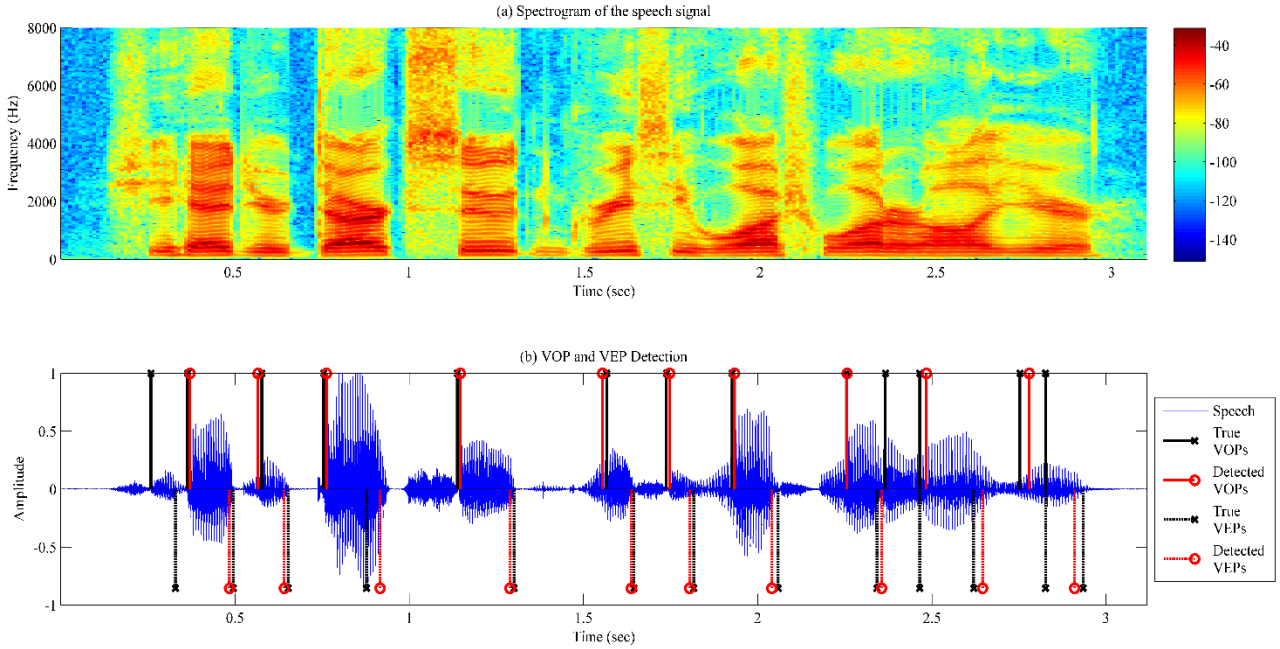

Fig. 5.7. Results of the VOP and VEP detection.

5.5 PERFORMANCE EVALUATION

The performance of the VOP and VEP detection method is evaluated on speech from 32 speakers in the TEST directory of the TIMIT corpus. Four speakers, consisting of two males and two females, are chosen from each of the 8 dialect regions. All 10 sentences from each speaker were used for testing, giving a total of 320 tests. The phoneme transcription files available in the TIMIT database contain the location of phone boundaries. The locations of the vowel boundaries from these files is used as the ground truth/reference. The performance of the VOP and VEP detection method is measured using the following parameters:

1. *Identification rate (IR):* A true VOP/VEP is said to have been identified if the proposed method has detected a VOP/VEP within ± 50 msec around it. The percentage of true VOPs/VEPs that are matched to a detected VOP/VEP within ± 50 msec is called the identification rate.
2. *Miss rate (MR):* The percentage of true VOPs/VEPs for which no detected VOP/VEP was found within ± 50 msec ($MR = 100\% - IR$).
3. *Spurious rate (SR):* The percentage of VOPs/VEPs detected by the proposed method, which are not within ± 50 msec of any true VOP/VEP.
4. *Root Mean Squared Deviation (RMSD):* The root mean square of the difference between the true and detected VOP/VEP locations (in msec) of all the VOPs/VEPs that were identified.

5.5.1 Clean Conditions

First the performance of the detection algorithm was tested in clean conditions, over the different sentence categories - SA, SI, and SX. The results are shown in Table 5.1.

Table 5.1: Performance of VOP and VEP detection in clean conditions.

Event	VOP				VEP			
Sentence Type	IR (%)	MR (%)	SR (%)	RMSD (ms)	IR (%)	MR (%)	SR (%)	RMSD (ms)
SA	81.07	18.93	13.88	16.82	79.00	21.00	16.08	19.83
SI	82.53	17.47	13.84	14.23	80.08	19.92	16.40	17.62
SX	82.14	17.86	15.57	15.98	79.75	20.25	18.02	17.74
All	82.05	17.95	14.66	15.61	79.70	20.30	17.10	18.15

Overall, the identification rate of VOPs was slightly higher than that of VEPs. This could be because slow variations at the end of a vowel are not as prominent as those at the onset, making the detection of VEPs more challenging. The spurious rate and the RMSD are also higher in the case of VEP detection. This suggests that the timing errors are larger for VEP detection. The performance of both detection algorithms was lowest for the SA sentences, which are meant to

expose dialect variations. The detection algorithm showed higher identification rates on the phonetically-compact (SX) sentences, which contained a good coverage of pairs of phones. However, the spurious rate and the RMSD were also higher for the SX sentences. This could be attributed to the extra occurrences of difficult phonetic contexts in the SX sentences. The best performance was obtained on the phonetically-diverse (SI) sentences.

5.5.2 Noisy Conditions

Next, we test the robustness of the VOP and VEP detection in the presence of different kinds of noise. The noise signals were obtained from the Signal Processing Information Base (SPIB) noise dataset [95]. The performance evaluation is conducted for clean speech, as well as for speech contaminated by white, pink, babble, vehicle, factory, and cockpit noise, at SNR levels of 30, 20, 10 and 0 dB. The results are presented in Table 5.2. In each test, the noisy speech signal was created by selecting a random segment from the noise file, scaling the noise to obtain the appropriate SNR, and then adding it to the clean speech signal.

Table 5.2: Performance evaluation of VOP/VEP detection in noise.

Event		**VOP**				**VEP**			
Noise Type	**SNR (dB)**	**IR (%)**	**MR (%)**	**SR (%)**	**RMSD (ms)**	**IR (%)**	**MR (%)**	**SR (%)**	**RMSD (ms)**
Clean		82.05	17.95	14.66	15.61	79.70	20.30	17.10	18.15
White	30	81.92	18.08	14.07	15.67	79.37	20.63	16.74	18.23
	20	81.37	18.63	13.23	15.61	78.64	21.36	16.13	18.38
	10	80.06	19.94	11.85	15.86	76.53	23.47	15.74	18.97
	0	75.49	24.51	15.78	17.59	71.15	28.85	20.62	20.43
Pink	30	81.87	18.13	14.26	15.63	79.43	20.57	16.82	18.24
	20	81.44	18.56	13.54	15.64	78.57	21.43	16.60	18.42
	10	79.70	20.30	12.66	15.82	76.10	23.90	16.61	18.89
	0	73.22	26.78	16.29	17.33	70.05	29.95	19.92	20.81
Babble	30	81.90	18.10	14.35	15.60	79.73	20.27	16.62	18.26
	20	81.32	18.68	14.09	15.60	78.85	21.15	16.70	18.39
	10	79.00	15.00	21.00	15.74	75.97	24.03	18.26	19.01
	0	73.73	26.27	27.75	18.27	70.75	29.25	30.66	21.36
Cockpit	30	81.80	18.20	14.50	15.56	79.53	20.47	16.87	18.28
	20	81.54	18.46	13.58	15.59	78.52	21.48	16.78	18.33
	10	80.11	19.89	12.59	15.78	76.60	23.40	16.42	18.92
	0	74.53	25.47	16.97	17.32	71.05	28.95	20.84	20.52
Factory	30	81.85	18.15	14.33	15.52	79.48	20.52	16.81	18.23
	20	81.19	18.81	13.67	15.56	78.49	21.51	16.54	18.33
	10	79.70	20.30	13.61	15.75	75.90	24.10	17.74	18.91
	0	73.45	26.55	25.38	17.99	70.40	29.60	28.48	21.13
Car	30	82.10	17.90	14.63	15.64	79.78	20.22	17.04	18.19
	20	81.92	18.08	14.72	15.67	79.65	20.35	17.09	18.18
	10	81.57	18.43	14.64	15.57	79.17	20.83	17.15	18.20
	0	80.86	19.14	14.32	15.62	77.99	22.01	17.37	18.43

From Table 5.2, we see that the VOP and VEP detection algorithms are quite robust against all noise types at SNR levels of 30 to 10 dB. A slight drop in identification rate is observed as the SNR is reduced from 30 to 10 dB. However, there is no sharp increase in the spurious rate or in the RMSD. In fact, a slight drop in the spurious rate is observed. The detection algorithms seem to be worst affected by babble noise. In particular, at 10 dB SNR, a sharp increase in the spurious rate is observed in both VOP and VEP detection under babble noise.

At 0 dB SNR, the identification rates drop significantly for all noise types except vehicle noise. An increase in the spurious rate and the RMSD is also observed in all the cases. The performance seems to be affected most by babble noise, followed by factory noise - high spurious rates are observed in these cases.

Thus, we have implemented a noise-robust method for detecting the Vowel Onset Points (VOP) and Vowel End Points (VEP) in a speech signal. In the next chapter, we study the speaker-discriminative power of transitions into and out of vowels using this method.

6 SPEAKER IDENTIFICATION EXPERIMENTS

In this chapter, the set-up and performance evaluation of our baseline speaker identification system using LSF features is explained. Subsequently, the relative importance of three zones of speech, namely vowel, non-vowel, and transition are investigated. Whether utilizing the speaker-discriminative zones of the speech signal and discarding less important zones during testing improves the recognition performance is also explored. In order to improve the robustness of the system, fusion of static and dynamic information obtained from the LSF features is examined.

6.1 PRELIMINARY EXPERIMENTS

In order to set up our LSF based speaker identification (SI) system, the effect of a number of parameters on the identification performance needs to be analyzed:

1. Dimension of the feature vector (LSF/LP order *p*)
2. Number of GMM mixture components (*M*)
3. GMM training procedure
4. Training and test utterance lengths

For this purpose, a number of preliminary experiments were conducted on a set of 32 speakers from the TEST directory of the TIMIT corpus. Four speakers, consisting of two males and two females, are chosen randomly from each of the 8 dialect regions. The training set for each speaker consists of the 5 SX and the 3 SI sentences. The test set consists of the two remaining SA sentences.

6.1.1.1 Enrollment

For the enrollment of each speaker, the 8 sentences in the training set are normalized by their maximum amplitude and concatenated. The concatenated speech is divided into 20 msec long frames, with a 10 msec frame shift. Thus, a training utterance of *n* seconds corresponds to $\Gamma = 100n$ frames. Among the Γ frames, LSF feature vectors of order *p* were extracted only from those frames classified as speech by the VAD. These LSF features were used for training a GMM with *M* mixture components. The GMM training procedure was initialized using the k-means clustering algorithm. As explained in Chapter 4, nodal, diagonal covariance matrices were used.

6.1.1.2 Evaluation

The performance evaluation was conducted in the following manner. For each speaker, the 2 sentences in the test set are normalized and then concatenated. The concatenated test speech is broken down into 20 msec long frames, with a 10 msec frame shift. A test utterance of *n* seconds corresponds to $\Gamma = 100n$ frames. The sequence of test frames was divided into overlapping segments of Γ frames each. Every segment of Γ frames is used as a separate test. In each Γ frame

segment, LSF features of order p were extracted from only the speech frames. These features were used for scoring across all speaker models. The speaker whose GMM produced the highest similarity score is the identified speaker. The identified speaker of each test segment is compared with the true speaker of the test utterance. The identification accuracy is the percentage of correctly identified test segments over all possible test segments.

6.1.2 LSF Order and Number of Components

A simple experiment was conducted to determine the optimal order p of the LSF feature vector and the number M of mixture components. The identification accuracy of the 32 speaker system described above was determined for different LSF orders, $p = 8, 12, 16, 20, 24, 28$ and for different numbers of mixture components per GMM, $M = 2, 4, 8, 16, 32, 64$. The entire training set was used for enrollment and a 1 second test utterance length was used. The results are shown in Fig. 6.1.

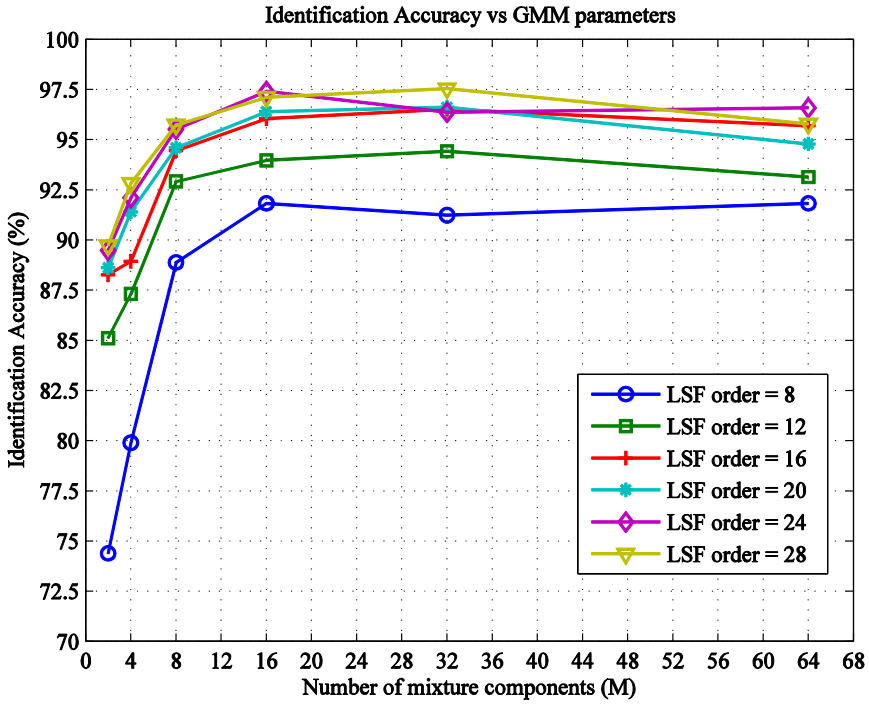

Fig. 6.1: Effect of GMM parameters on identification accuracy.

From Fig. 6.1, we see that the identification accuracy increases sharply as the number of mixture components is increased from 2 to 8, leveling off at M=16. As the value of M is increased further, no substantial improvement in identification accuracy is observed. Therefore, the number of mixture components, $\boldsymbol{M}$**=16** is selected for our speaker identification system.

The choice of prediction order depends on the sampling frequency of the speech signal. A sampling frequency of 16 kHz corresponds to a speech bandwidth of 8 kHz. For average adult speech, formants are spaced so that there is approximately 1 formant per 1 kHz. Thus, for an 8 kHz speech bandwidth, there are approximately 8 formants. Since each formant corresponds to a pair of

complex-conjugate poles, a minimum LP order of 16 is required. It is recommend to use 4 or 5 additional coefficients to model the spectral slope, closely spaced formants, and the effect of anti-resonances on the spectrum [78]. Thus, it appears that an optimal LSF order would be 20.

To verify our hypothesis, we observe the effect of LSF order on the identification accuracy in Fig. 6.1. As the LSF order increases from 8 to 16, we see a substantial improvement in the identification accuracy. However, as the LSF order is increased further, the increase in identification accuracy is quite small. This matches with our previous reasoning, and hence an LSF order of $\boldsymbol{p}$**=20** is selected for our speaker identification system.

6.1.3 Initialization of GMM Training

As explained in Chapter 4, k-means clustering is used for the initialization of the EM algorithm during GMM training. However, the k-means clustering algorithm is initialized with a random starting point. Hence, we expect the GMM training procedure to converge to different local maxima depending on the initial condition. To observe this variation, we repeat the GMM training procedure (k-means initialization, followed by EM algorithm) ten times for one speaker, on the same training feature set. The LSF order is chosen to be 20, and 16 mixture components are used. The log-likelihoods of the GMMs obtained in each run are shown in the box plot in Fig. 6.2(a).

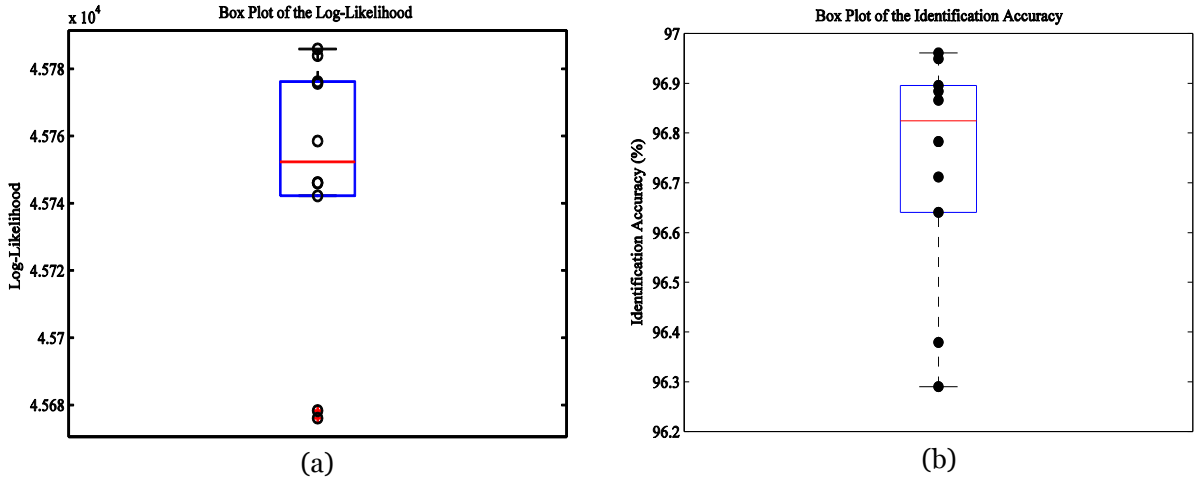

Fig. 6.2: (a) Box plot of log-likelihood from 10 runs of the GMM training procedure. (b) Effect of initialization of GMM training on identification accuracy (test utterance length: 1 sec).

From the box plot, we observe that the random initialization of the k-means clustering has an effect on the log-likelihood of the resulting GMM. However, the distribution of the log-likelihoods is skewed towards the top, and we observe that a number of initializations result in nearly the same likelihood values. This suggests that the best-fit GMM for each speaker can be found by repeating the GMM training procedure a sufficient number of times. In our baseline system, the GMM training procedure is repeated 10 times for each speaker, resulting in 10 GMMs. Among these GMMs, the one with the highest log-likelihood is chosen to be representative of that speaker.

To observe the effect of model initialization on the speaker identification performance, the entire enrollment and evaluation procedure is repeated 10 times. The LSF order is chosen to be 20, and 16 mixture components are used. In each run, a different set of GMMs is obtained for the 32 speakers, and the identification accuracy is evaluated for a 1 second test utterance length. The results are shown in the box plot in Fig. 6.2(b). We observe that the model initialization does have an effect, albeit minor, on the identification accuracy. The distribution of the identification accuracy seems to be skewed towards the top and the variation is quite small, on the order of 0.6%.

6.1.4 Training vs Test Utterance Lengths

Next, the effect of the amount of training data on the performance of the speaker identification system is investigated. Speaker models were trained using different training utterance lengths, ranging from 3 seconds to 18 seconds. The LSF order is chosen to be 20, and 16 mixture components are used. The test utterance length is maintained at 1 second. The identification accuracy versus the length of training utterance is plotted in Fig. 6.3. As expected, identification accuracy increases with an increase in the amount of training data. The best identification performance is obtained for more than 15 seconds of training data.

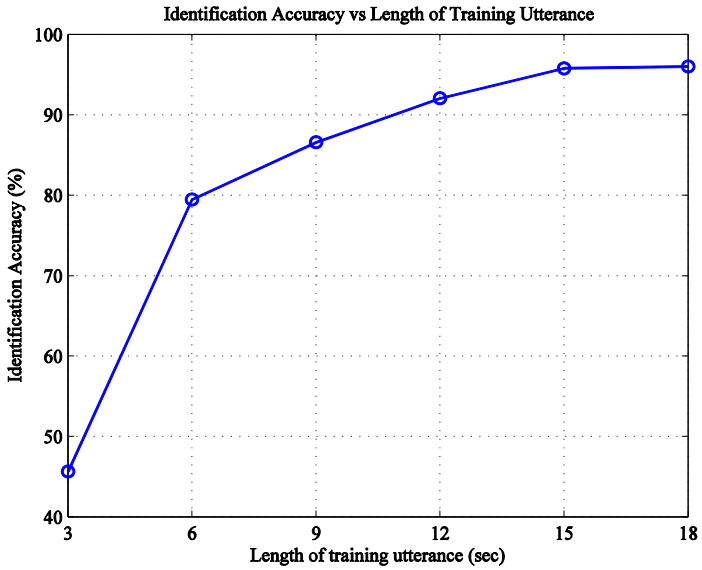

Fig. 6.3: Identification accuracy vs length of training utterance

In order to investigate the effect of the test utterance length on the performance of the speaker identification system, the following experiment is conducted. Speaker models are trained using the entire training speech, and evaluated for various test utterance lengths ranging from 0.5 seconds to 4 seconds. The results are shown in Fig. 6.4. We observe that the identification accuracy increases with the length of the test utterance. The identification accuracy reaches 100% and levels off for test utterance lengths greater than 3 seconds.

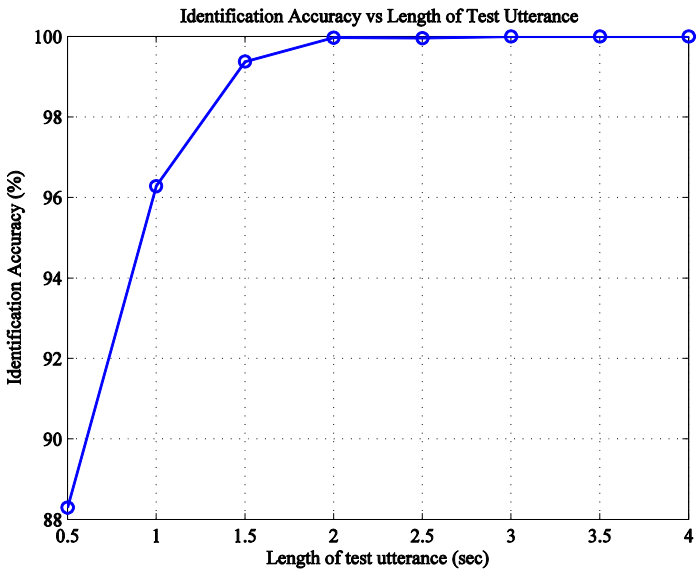

Fig. 6.4: Identification accuracy vs length of the test utterance.

6.2 EXPERIMENTAL SETUP

Based on the results of the preliminary experiments, the speaker identification experiments in the sequel are set up in the following manner.

6.2.1 Speaker Set

The set of 168 speakers in the TEST directory of the TIMIT corpus are selected for the training and evaluation of our speaker identification system. This set consists of 112 male speakers and 56 female speakers. The training set for each speaker consists of the 5 SX and the 3 SI sentences. All the sentences in the training set are normalized and concatenated to produce approximately 24 seconds of training speech. The test set consists of the two SA sentences.

6.2.2 Speaker Enrollment

For each speaker, a Gaussian Mixture Model is created using the following procedure. The training speech is divided in 20 msec long frames, with a frame shift of 10 msec. LSF feature vectors of order p=20 are extracted from the speech frames and used for training. The EM algorithm, initialized using k-means clustering is used to train a GMM with M=16 mixture components using the training features. Nodal, diagonal covariance matrices were used. For each speaker, the GMM training is repeated 10 times, and the GMM with the highest log-likelihood is chosen to be the final model for that speaker.

6.2.3 Performance Evaluation

6.2.3.1 Clean Conditions

The performance of the LSF based speaker identification system is evaluated under clean conditions as follows. The SA sentences in the test set are used as two individual tests, of approximately 3 seconds each. Thus, a total of $168 \times 2 = 336$ tests are conducted. Each test utterance is divided in 20 msec long frames, with a frame shift of 10 msec. LSF feature vectors of dimension p=20 are extracted from the speech frames and used for evaluation. These features were used for scoring across all speaker models. The speaker whose GMM produced the highest similarity score is taken as the identified speaker. The identified speaker in each test is compared with the true speaker. The identification accuracy is the percentage of correctly identified speakers across all 336 tests.

6.2.3.2 Noisy Conditions

The performance evaluation is conducted for speech contaminated by white, pink, babble, cockpit, factory, and car noise, at SNRs from 30 dB to 0 dB, in steps of 6 dB. For a particular noise type at a particular SNR, the evaluation is conducted in the following manner. For every speaker, each of the two SA test sentences is corrupted with 5 random realizations of noise. The noisy speech signals were created by selecting a random segment from the noise file, scaling the noise segment to obtain the appropriate SNR and adding it to the clean speech signal. The scaling factor is given by $a = \sqrt{P_s / (SNR \times P_n)}$, where P_s and P_n are the speech and noise segment powers respectively. Thus, a total of $168 \times 2 \times 5 = 1680$ tests are conducted at each SNR, for each noise type. The identification accuracy is the percentage of correctly identified speakers across all 1680 tests.

6.3 SPEAKER IDENTIFICATION USING LSF FEATURES

The parameters of our baseline speaker identification system are outlined in the following table.

Table 6.1: Parameters of the LSF based SI system.

Parameter	**Description**
Number of speakers (S)	168
Training set of each speaker	All SX, SI sentences (~3 seconds x 8)
Test set of each speaker	SA sentences (~3 seconds x 2)
Feature Type	LSF
Feature Dimension/Order (p)	20
Frame Length (L)	20 msec
Frame Shift (δ)	10 msec
Number of GMM Components (M)	16
GMM Covariance Type	Nodal and Diagonal

6.3.1 Performance Evaluation

In this section, the performance evaluation of the LSF based speaker identification system is presented. For comparison, the performance of a speaker identification system that uses MFCC features is also evaluated. The pre-processing and frame blocking was performed in the same manner as that for LSF feature extraction. A total of 21 MFCC coefficients were derived using 24 filters applied to the magnitude spectrum of 20 msec long frames, with a frame shift of 10 msec. The resulting MFCC coefficients were liftered using a liftering parameter of 22. The first MFCC coefficient was discarded, and the 20 other coefficients were considered as the MFCC feature vector for each frame. The rest of the training and evaluation process remains the same.

Under clean conditions, the LSF based system has an identification accuracy of 99.7%. On the other hand, an identification accuracy of 99.1% was obtained using MFCC coefficients. The results show that the LSF based speaker identification system performs comparably to an MFCC based speaker identification system for the selected configuration.

Table 6.2: Performance of the LSF based SI system under clean conditions.

Feature Type	**LSF**	**MFCC**
Identification Accuracy (%)	99.70	99.10

Next, the performance of the LSF based speaker identification system is evaluated in the presence of different types of background noise, as described in Section 6.2.3.2. The results are presented in Table 6.3, and also represented in Fig. 6.5.

Table 6.3: Performance of the LSF based SI system under noisy conditions.

SNR (dB)	**Identification Accuracy (%)**						
	White	**Pink**	**Babble**	**Cockpit**	**Factory**	**Car**	**Average**
30	56.49	95.54	99.70	99.05	99.29	99.70	91.63
24	15.48	65.18	99.64	93.21	95.89	99.70	78.18
18	4.46	20.36	99.29	54.82	68.51	99.70	57.86
12	1.55	2.02	83.04	12.86	24.70	99.70	37.31
6	1.01	1.01	40.00	3.93	4.05	99.64	24.94
0	0.83	0.60	8.87	1.55	0.77	97.14	18.29
Clean	99.70						

From Fig. 6.5, we observe that the performance of the LSF based SI system degrades considerably in the presence of most noise types. The SI system seems to be most immune to car noise, for which the identification accuracy begins to drop only at 0 dB SNR. The SI system is also reasonably robust against babble noise. The identification accuracy is worst affected by pink noise and white noise. With pink noise, at 0 dB SNR, the performance is observed to fall to 0.5952% which is equivalent to making a random guess. The SI system is highly sensitive to white noise - even at a high SNR of 30 dB, where the identification accuracy already dropped sharply, to 56.4%.

This might be because white and pink noise affect the entire frequency spectrum, whereas the other noise types are mostly low-frequency in nature.

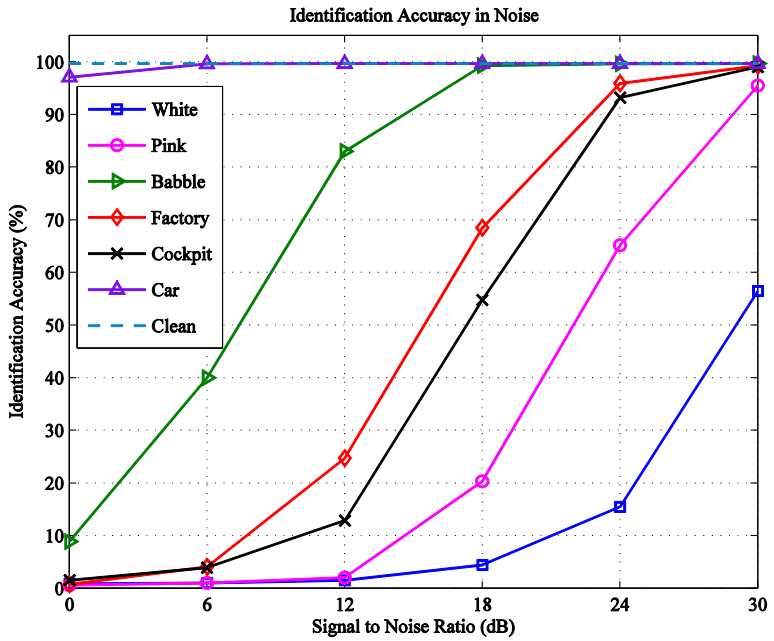

Fig. 6.5: Performance of the LSF based SI system under background noise.

Thus, we conclude that while the LSF based SI system provides near perfect performance under clean conditions, the performance degrades considerably under most types of noise. Next, we investigate whether there is any benefit in selectively utilizing certain speech frames which contain the most speaker-specific information during identification, as opposed to utilizing frames from the entire utterance. In particular, we wish to investigate whether transitions into and out of vowels contain speaker-specific information that makes the LSF based SI system more robust in noisy environments.

6.3.2 Discriminative Power of Speech Zones

For the LSF based SI system, the discriminative power of different speech zones is analyzed, under clean as well as noisy conditions. This analysis is conducted during the evaluation phase in the following manner. In each test utterance, the Vowel Onset Points (VOPs) and Vowel End Point (VEPs) are detected using the noise-robust detection algorithm presented in Chapter 5. Based on the VOP and VEP locations, each speech frame in the test utterance is classified as either a vowel, non-vowel, or transition frame. Here, 'transition' refers only to transitions into and out of vowels (CV/VC boundaries). This classification process is illustrated in Fig. 6.6.

A VOP marks the end of a consonant and the beginning of a vowel. Around 10-20 msec after the VOP, the steady vowel region begins. The dynamics of the speech spectrum vary rapidly around 20 msec before and after a VOP, as seen in the dynamic behavior of several of the LSF.

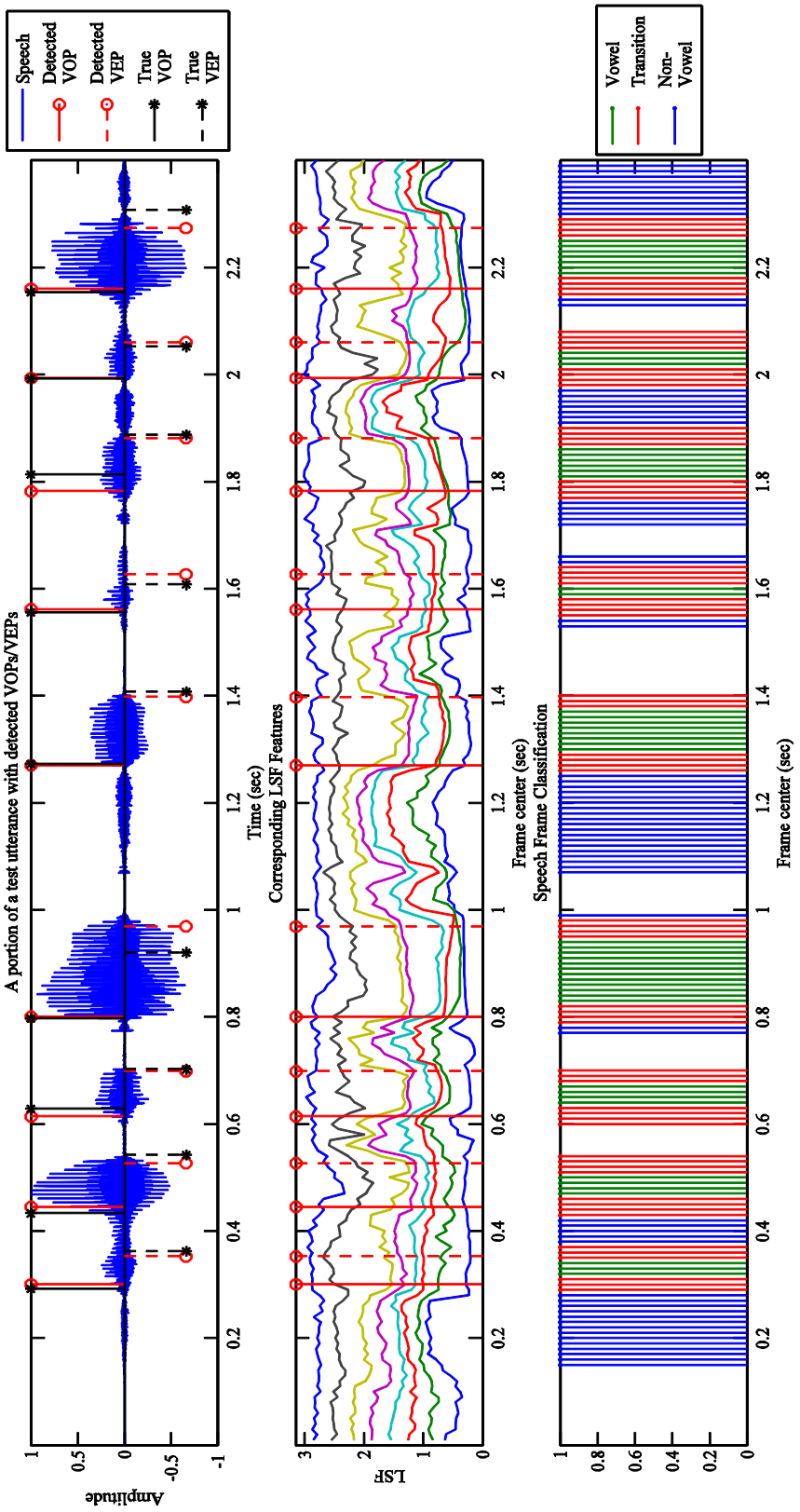

Fig. 6.6: Classification of speech frames.

Thus, frames whose centers lie within this window are marked as transition frames. For short vowels less than 50 msec long, frames whose centers lie within 20 msec before and 10 msec after the VOP are considered to be transition frames. Similarly, a VEP marks the end of a vowel and the beginning of a consonant. Speech frames whose centers lie within 20 msec before/after every VEP are marked as transition frames. For short vowels less than 50 msec long, frames whose centers lie within 10 msec before and 20 msec after the VEP are considered to be transition frames. The rest of the speech frames between each VOP and its corresponding VEP are marked as (steady) vowel frames. All remaining speech frames are marked as non-vowel frames. Non-speech frames are not considered for frame classification.

Next, the identification accuracy of the SI system is determined by scoring only on features extracted from transition frames, vowel frames, and non-vowel frames respectively. This is equivalent to giving a weight of 1 to only the frames of interest, and allocating a zero weight to all other frames. For example, let X_{tr} denote the set of features obtained from frames marked as transitions. Using (4.29), the similarity score with the GMM of speaker *s*, obtained by considering only transition frames is:

$$\Phi_{s,X_{tr}} = \sum_{\mathbf{x} \in X_{tr}} \log p(\mathbf{x} \mid \lambda_s) \tag{6.1}$$

Then, the identified speaker by scoring only on transition frames is:

$$\hat{s}_{X_{tr}} = \arg \max_{1 \le s \le S} \Phi_{s,X_{tr}} \tag{6.2}$$

The category which is most speaker-discriminative is expected to produce the highest identification accuracy. The results are shown in Table 6.4, and are also represented in Fig. 6.7. Any category is considered to have outperformed another only if its performance is at least $(1/1680) \times 100 = 0.06\%$ higher under noisy conditions, and at least $(1/336) \times 100 = 0.30\%$ higher in clean conditions. In each row of Table 6.4, the best performance among the four categories is underlined. Results which outperform the baseline (all frames) are highlighted in bold.

Under clean conditions, scoring selectively on some frames provided a slightly lower performance compared to scoring on all speech frames. Thus, it is better to utilize all of the available frames, rather than perform frame-level selection. Observe that, among the different categories, the transition frames are the most speaker-discriminative. An identification accuracy of 99.4% was obtained by scoring only on transition frames, which is nearly the same as that obtained by scoring on all frames (99.7%). Vowel frames came in as a close second, with an identification accuracy of 97.6%. Combining information from transition and vowel frames did not improve the SI performance.

These results suggest that in clean conditions, rapidly varying transition regions are more useful compared to steady vowel regions, which contain more static and redundant information. This result confirms our hypothesis that transitions contain the most speaker-specific information.

Table 6.4: Analysis of speaker discriminative power of different speech zones in an LSF based SI system.

Noise Type	SNR (dB)	Identification Accuracy (%) of the LSF based SI system				
		All (Baseline)	Non-Vowel	Transition	Vowel	Vowel + Transition
Clean		99.70	91.07	99.40	97.62	99.10
White	30	56.49	28.81	44.35	54.05	54.35
	24	15.48	7.14	12.92	**20.60**	**18.75**
	18	4.46	3.45	3.57	3.75	3.21
	12	1.55	**1.85**	1.43	1.07	1.19
	6	1.01	0.48	**1.25**	0.89	**1.55**
	0	0.83	0.48	0.71	0.29	**0.89**
Pink	30	95.54	66.79	88.81	90.89	93.39
	24	65.18	32.20	53.33	63.45	64.64
	18	20.35	5.12	19.76	**25.54**	**23.39**
	12	2.02	1.61	**2.56**	**5.54**	**3.39**
	6	1.01	**1.13**	0.65	0.71	0.71
	0	0.60	**0.95**	0.60	**0.83**	0.60
Babble	30	99.70	89.64	99.46	97.62	99.05
	24	99.64	85.12	99.23	97.74	98.93
	18	99.29	62.74	98.10	97.08	98.69
	12	83.04	19.70	**87.80**	**91.01**	**92.74**
	6	40.00	3.81	**48.63**	**65.65**	**63.57**
	0	8.87	1.49	**12.68**	**23.75**	**18.15**
Cockpit	30	99.05	81.01	96.85	97.74	98.39
	24	93.21	59.17	87.56	88.27	91.37
	18	54.82	15.89	49.70	**61.73**	**60.71**
	12	12.86	3.27	**13.63**	**21.25**	**19.23**
	6	3.93	2.32	**4.40**	**7.20**	**6.25**
	0	1.55	1.25	1.37	**2.02**	**1.61**
Factory	30	99.29	83.63	98.39	97.32	98.33
	24	95.89	63.69	92.08	91.73	94.23
	18	68.51	21.90	62.14	**71.55**	**71.85**
	12	24.70	2.80	**28.10**	**39.23**	**35.48**
	6	4.05	0.89	**5.48**	**11.31**	**8.69**
	0	0.77	**0.83**	**0.89**	**2.50**	**1.55**
Car	30	99.70	91.49	99.40	97.62	99.11
	24	99.70	90.95	99.40	97.50	98.99
	18	99.70	89.52	99.40	97.32	98.93
	12	99.70	84.88	99.40	97.02	98.93
	6	99.64	74.05	99.11	96.90	98.87
	0	97.14	50.77	**97.86**	96.37	**98.27**

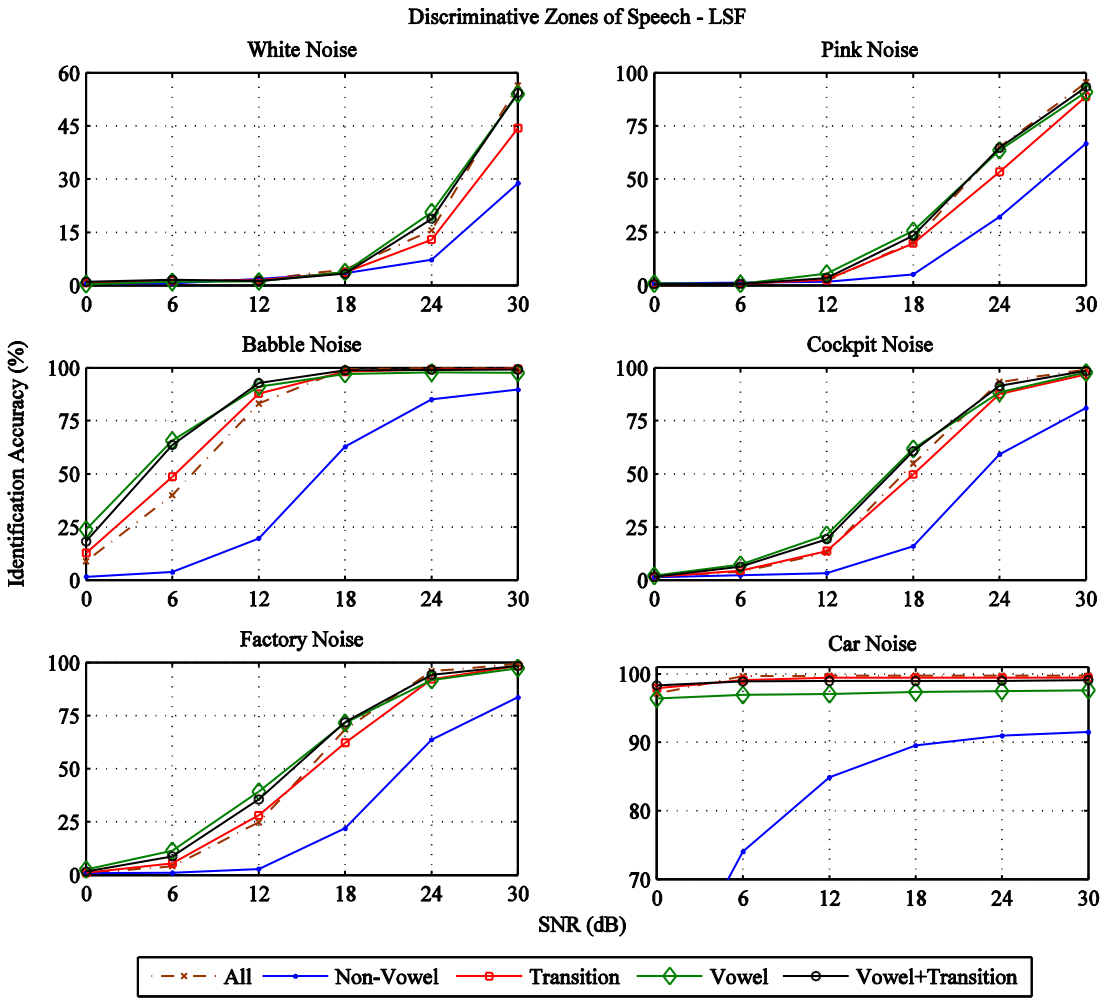

Fig. 6.7: Discriminative power of different speech zones under noisy conditions.

By scoring on non-vowel regions only, an identification accuracy of only 91% was achieved. This might be attributed to the lower energy of non-vowel regions. In noisy conditions, non-vowel frames appear to be the least robust and least speaker-discriminative. This result is expected, since non-vowel frames are low in energy and are therefore most heavily affected by noise. In addition, at low SNRs, the energy based VAD begins to misclassify many non-speech frames as speech frames. These misclassified frames are in turn classified as non-vowel frames by the frame classification algorithm. This could be another reason for the sharp decline in SI performance based on non-vowel frames.

Next, the SI performance based on vowel and transition frames is compared under different kinds of noise. As observed earlier, the SI system is worst affected by white noise. Transition regions seem to be sensitive to white noise, as scoring only on transition frames almost always performs much lower than the baseline. At 30 dB SNR, frame-level selection did not result in any performance improvement. However, at 24 dB SNR, scoring only on vowel frames outperforms the baseline. Therefore, vowel regions appear to be more robust at lower SNR. From below 18 dB,

the performance of the system is severely affected. Hence, it is no longer possible to observe any clear differences between the different categories.

A similar result is observed under pink noise. Up to 24 dB SNR, frame-level selection does not produce any performance improvement over the baseline. Transition frames seem to be strongly affected by pink noise, while vowel regions appear to be more immune. However, combining transition frames with vowel frames boosts the identification accuracy. At SNRs below 18 dB, frame-level selection begins to result in performance improvements. Scoring only on vowel frames leads to the best identification accuracy, outperforming the baseline considerably. From below 6 dB, the performance of the system is severely affected. No clear difference in SI robustness associated with using different speech zones is observed.

The SI system is much more immune to babble noise. From 30 dB to 18 dB SNR, the best performance is obtained by scoring on all speech frames. At these SNR levels, transition frames are observed to be the most speaker-discriminative. From 12 dB SNR and below, scoring on all frames is observed not to be the best approach. Although scoring on transition frames only does improve performance, scoring only on vowel frames results in the highest performance gain. From 6 dB and below, even when scoring on transition and vowel frames together, the performance was lower than when scoring exclusively on vowel frames. Thus, we conclude that at lower SNRs, though transition frames do contain speaker-specific information, they are outdone by vowel frames.

Similarly, under cockpit and factory noise, at SNR levels of 30 dB - 24 dB, the best performance is obtained by scoring on all frames. In addition, among the different categories, transition and vowel frames are observed to be the most speaker-discriminative. From 18 dB SNR and below, frame-level selection boosts the identification accuracy. At these SNR levels, transition frames begin to lose their speaker-discriminative power. Scoring on transition frames does improve performance, but scoring on vowel frames only outperforms the baseline by a considerably bigger margin.

Under car noise, a different behavior is observed since the SI system is extremely robust against this noise type. In this case, the transition frames are seen to be the most speaker-discriminative, and do not lose their power under noise. Scoring on transition frames only results in nearly the same SI performance as the baseline, even at 0 dB SNR.

6.3.3 Conclusions

From the above analysis, we come to a few major conclusions. Under low noise conditions, it is better to score on all speech frames instead of selectively scoring on certain frames. However, when noise levels are high, frame-level selection is able to provide significant performance gains.

For this LSF based SI system, transition zones are clearly the most speaker-discriminative in clean and high SNR conditions. However, they are sensitive to noise. In fact, transition frames appear to be severely affected by white and pink noise. Scoring on transition frames only does not provide

any performance gain for these noise types. In other noise types, though transition frames are speaker-discriminative, they are outperformed by vowel frames in low SNR conditions.

The reason for this might be two-fold. Firstly, since steady vowels are high energy regions, they are expected to be most robust under noisy conditions. On the other hand, the information in the CV/VC transition zones might become buried in the noise. Also, since we are using only static LSF features, we might not be adequately capturing the information present in the transition regions. Thus, the next step would be to try and model the dynamic information present in the transition regions in a better way.

6.4 SPEAKER IDENTIFICATION USING DELTA-LSF FEATURES

Our baseline speaker identification system in the previous section utilized 'static' LSF feature vectors. Each LSF vector represents the spectral content of a short-term frame in the speech signal. There is no temporal information in these LSF features. In order to incorporate more dynamic/transitional information, first-order spectro-temporal features known as delta-LSF (ΔLSF) features are utilized.

6.4.1 Delta-LSF Features

There are two general ways in which ΔLSF features are calculated – differentiation or linear regression. Although the differentiation method is simple to implement, it acts as a high-pass filtering operation in the LSF domain. This in turn might amplify noise and lower the identification accuracy, especially in noisy conditions. Thus, the linear regression method is chosen to compute the ΔLSF features. In this technique, an orthogonal polynomial curve is fitted to each LSF coefficient trajectory, over a finite window of length $2K + 1$ frames [34]. The 1st-order orthogonal polynomial coefficient, or the generalized spectral slope (in time) is then denoted by:

$$\Delta \mathbf{x}_t = \frac{\sum_{k=-K}^{K} k\mathbf{x}_{t+k}}{\sum_{k=-K}^{K} k^2} \tag{6.3}$$

Here, $\Delta \mathbf{x}_t$ refers to the t^{th} ΔLSF feature vector, and $\mathbf{x}_t$ refers to the t^{th} LSF feature vector. Higher-order polynomials can be used to obtain smoother estimates, but in practice the first order polynomial is shown to be adequate. The value of K is typically chosen from the 1 to 3 range. Since the intent is to capture the dynamic information without too much smoothing, a value of K=2 is selected. Before computing the ΔLSF features, the LSF features are padded with K extra replicate frames at both ends.

6.4.2 Performance Evaluation

In this section, the effect on the identification performance of using ΔLSF features instead of static LSF features is analyzed. In order to do this, a speaker identification system is set up and evaluated using the same parameters as in Table 6.1. The only difference is the use of ΔLSF features of order p=20, instead of LSF features, for training and evaluation. The results of the evaluation of SI performance based on ΔLSF features are presented in Table 6.5. A comparison between the performance based on LSF and based on ΔLSF features under noise is shown in Table 6.6 and Fig. 6.8.

Table 6.5: SI Performance of the delta-LSF based SI system in noisy conditions.

SNR (dB)	Identification Accuracy (%) of the ΔLSF based SI system						
	White	Pink	Babble	Cockpit	Factory	Car	Average
30	35.95	77.74	93.99	90.77	91.73	92.98	80.53
24	10.65	40.48	92.14	76.90	80.18	92.56	65.49
18	1.73	10.06	87.32	33.45	46.73	91.79	45.18
12	0.65	1.07	64.46	7.56	11.96	90.89	29.43
6	0.60	0.60	22.56	1.43	1.07	89.58	19.31
0	0.60	0.60	3.93	0.65	0.60	83.10	14.91
Clean	93.45						

Table 6.6: Performance comparison of LSF vs delta-LSF based SI systems.

SNR (dB)	Average Identification Accuracy (%)	
	LSF	ΔLSF
30	91.63	80.53
24	78.18	65.49
18	57.86	45.18
12	37.31	29.43
6	24.94	19.31
0	18.29	14.91
Clean	99.70	93.45

The performance obtained by using ΔLSF features is observed to be consistently worse than that using LSF features, under clean as well as noisy conditions. This is not an entirely unexpected result, since the computation of ΔLSF features is somewhat like differentiation, which amplifies noise. Under background noise, the behavior of the ΔLSF based SI system is similar to that of the LSF based SI system. The performance seems to be most immune to car noise, followed by babble noise. The identification accuracy is worst affected by pink and white noise. Under these noise types, the performance is observed to fall to below 1% at an SNR of 12 dB and lower. The SI system is highly sensitive to white noise - even at a high SNR of 30 dB, the identification accuracy is only 35.9%.

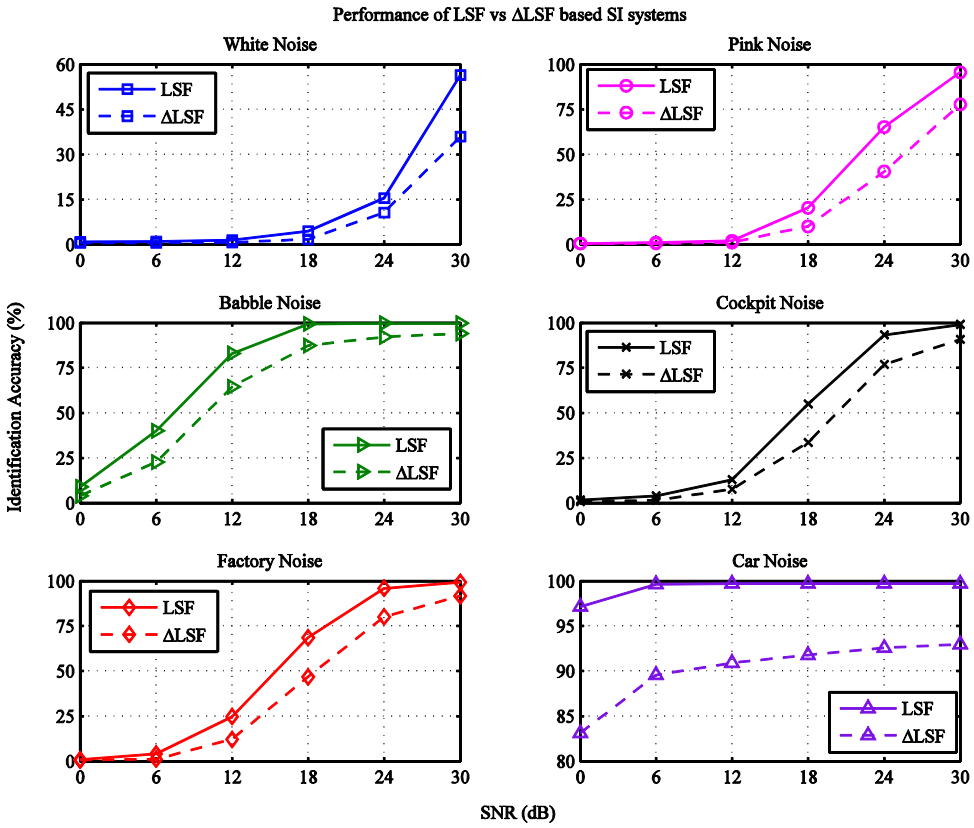

Fig. 6.8: Performance comparison of LSF vs delta-LSF based SI system.

The conclusion is that ΔLSF features are not ideal for speaker identification when used on their own. However, we suspect that a ΔLSF based SI system operates quite differently from an LSF based system. In order to test this hypothesis, the experiment that was conducted in Section 6.3.2 is now repeated for ΔLSF features. Since ΔLSF features capture dynamic information, the hypothesis is that an SI system which uses these features would find the transition zones of speech to be more useful or discriminative, compared to the steady vowel zones.

6.4.3 Discriminative Power of Speech Zones

Using the ΔLSF based SI system, we analyze the speaker-discriminative power associated with different speech zones, under clean as well as noisy conditions. The results are shown in Table 6.7, and also in Fig. 6.9.

Again, we infer that in clean and high SNR conditions, frame-level selection is not advantageous. It is more effective to utilize all available frames. However, frame-level selection is beneficial under lower SNR conditions. In clean as well as noisy conditions, non-vowel frames appear to be the least speaker-discriminative. Scoring on transition frames only results in higher identification accuracy than scoring on vowel frames only. This suggests that the transition regions are more speaker discriminative than vowel regions for a SI system which uses ΔLSF features. This result confirms our earlier line of reasoning.

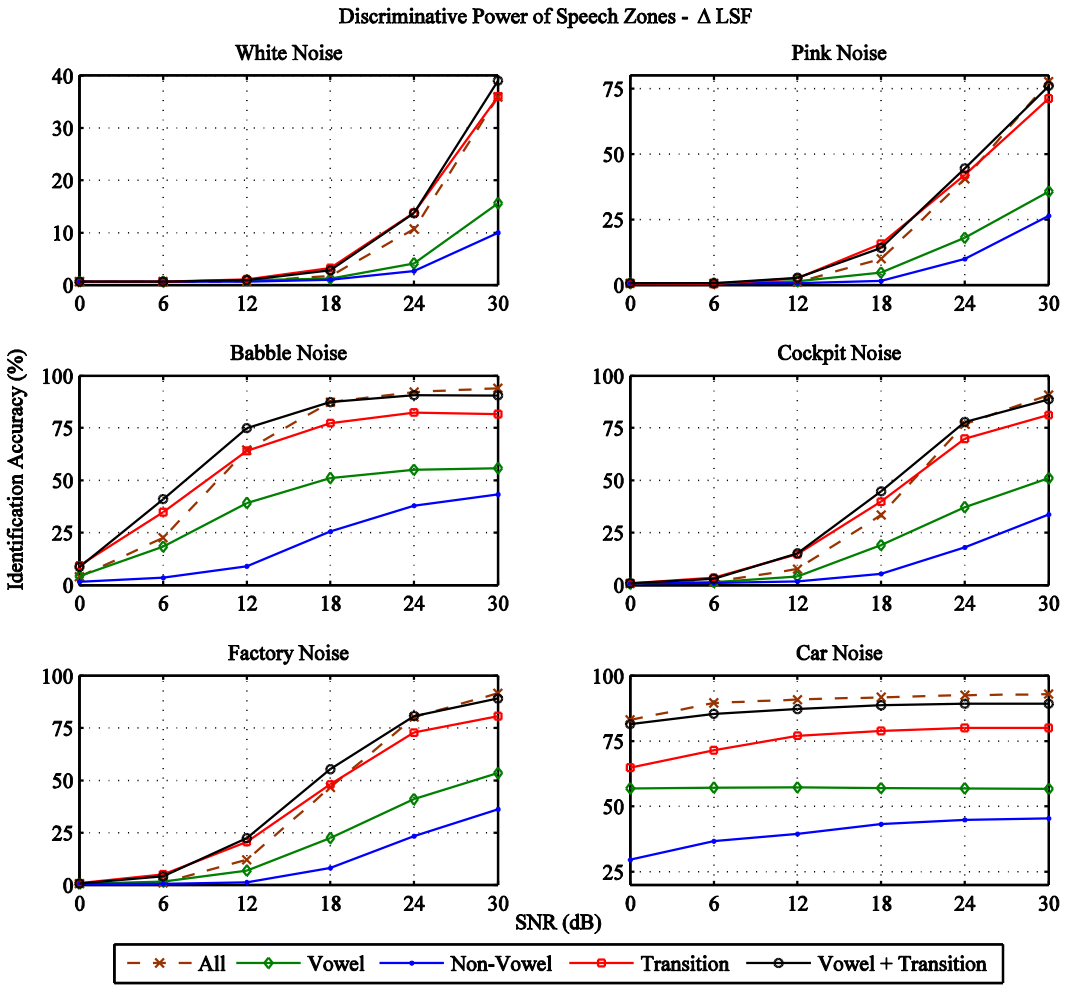

Fig. 6.9: Discriminative power of speech zones for the delta-LSF based SI system.

In fact, under low SNR conditions, better performance is achieved by scoring on transition frames only. However, the identification accuracy is improved even further by scoring on both vowel and transition frames combined. Thus, while vowel regions do not perform well on their own, they provide some important complementary information.

From the above analysis, the conclusion is that for a speaker identification system that uses static LSF features, transition frames are very speaker discriminative in cleaner conditions, but vowel regions are more speaker discriminative under noise. When dynamic LSF features are used, transition regions are the most speaker discriminative, but vowel regions contain some important complementary information.

Table 6.7: Analysis of speaker discriminative power of different speech zones in a delta-LSF based SI system.

Noise Type	SNR (dB)	Identification Accuracy (%) of the ΔLSF based SI system				
		All	Non-Vowel	Transition	Vowel	Vowel + Transition
Clean		93.45	47.92	81.55	56.55	89.58
White	30	35.95	10.00	**36.01**	15.60	**39.05**
	24	10.65	2.62	**13.81**	4.11	**13.69**
	18	1.73	1.01	**3.21**	1.19	**3.21**
	12	0.65	0.60	**1.07**	**0.89**	**1.07**
	6	0.60	0.60	0.60	0.60	0.60
	0	0.60	0.60	0.60	0.60	0.60
Pink	30	77.74	26.37	71.25	35.65	76.13
	24	40.48	9.88	**42.08**	18.04	**44.52**
	18	10.06	1.55	**15.77**	4.64	**14.17**
	12	1.07	0.60	**2.50**	**1.43**	**2.74**
	6	0.60	0.60	0.30	0.60	0.60
	0	0.60	0.60	0.54	0.60	0.60
Babble	30	93.99	43.27	81.67	55.83	90.54
	24	92.14	37.92	82.44	55.06	90.71
	18	87.32	25.42	77.38	51.07	87.32
	12	64.46	8.93	64.05	39.17	**74.88**
	6	22.56	3.33	**34.82**	18.33	**41.01**
	0	3.93	1.49	**9.35**	**4.17**	**8.63**
Cockpit	30	90.77	33.75	81.19	51.01	88.75
	24	76.90	17.98	69.88	37.20	**77.86**
	18	33.45	5.42	**39.82**	19.05	**44.70**
	12	7.56	1.67	**14.58**	4.05	**15.12**
	6	1.43	0.89	**3.33**	1.31	**3.04**
	0	0.65	0.65	**0.71**	**0.71**	**0.77**
Factory	30	91.73	36.19	80.71	53.51	89.05
	24	80.18	23.39	72.86	40.95	80.65
	18	46.73	8.10	**47.98**	22.50	**55.36**
	12	11.96	1.25	**20.48**	6.90	**22.32**
	6	1.07	0.60	**5.18**	**1.73**	**4.17**
	0	0.60	0.60	**0.77**	0.60	0.65
Car	30	92.98	45.48	80.00	56.73	89.3452
	24	92.56	44.94	80.00	56.96	89.2857
	18	91.79	43.27	78.93	57.08	88.6905
	12	90.89	39.46	77.02	57.38	87.2619
	6	89.53	36.73	71.55	57.26	85.4762
	0	83.10	29.64	64.88	56.90	81.4881

Thus, it is clear that LSF and ΔLSF based classifiers operate in a complementary manner, extracting information from different zones of speech. Intuitively, if two classifiers operate in a similar manner, i.e. if they misclassify the same speech segments, done cannot expect an improvement when they are combined. However, since these classifiers operate in a complementary manner, the next logical step would be to perform information fusion.

6.5 FUSION OF INFORMATION FROM LSF AND DELTA-LSF FEATURES

A common technique used to improve the accuracy and robustness of SIV systems is the fusion of static features with the corresponding dynamic features. There are two commonly used fusion techniques, namely feature-level fusion and score-level fusion.

6.5.1 Feature-level Fusion

In feature-level fusion (FLF), the ΔLSF features are fused with the LSF features at the feature vector level. First, both the LSF and the ΔLSF feature vectors (of order p=20) are normalized by their respective Euclidean norms. Then, these normalized vectors are concatenated to form a new feature vector of dimension 40. These LSF+ΔLSF feature vectors are then used for training and testing of an SI system with the same parameters as the static LSF based SI system. The results are shown in Table 6.8 and Table 6.9.

Table 6.8: Identification accuracy by feature-level fusion of LSF and delta-LSF.

SNR (dB)	Identification Accuracy (%) of Feature-level Fusion						
	White	Pink	Babble	Cockpit	Factory	Car	Average
30	48.27	91.13	99.52	97.92	98.21	99.70	89.13
24	12.44	53.63	99.29	83.57	89.70	99.58	73.04
18	2.86	10.54	96.67	40.30	54.70	99.52	50.76
12	1.13	1.01	71.25	8.63	13.21	99.40	32.44
6	0.42	0.60	26.85	1.37	2.50	98.63	21.73
0	0.48	0.60	6.31	1.01	1.01	92.26	16.94
Clean	99.70						

Table 6.9: Performance comparison of feature-level fusion with LSF and delta-LSF based SI systems.

SNR (dB)	Average Identification Accuracy (%)		
	LSF	ΔLSF	LSF+ ΔLSF (FLF)
30	91.63	80.53	89.13
24	78.18	65.49	73.04
18	57.86	45.18	50.76
12	37.31	29.43	32.44
6	24.94	19.31	21.73
0	18.29	14.91	16.94
Clean	99.70	93.45	99.70

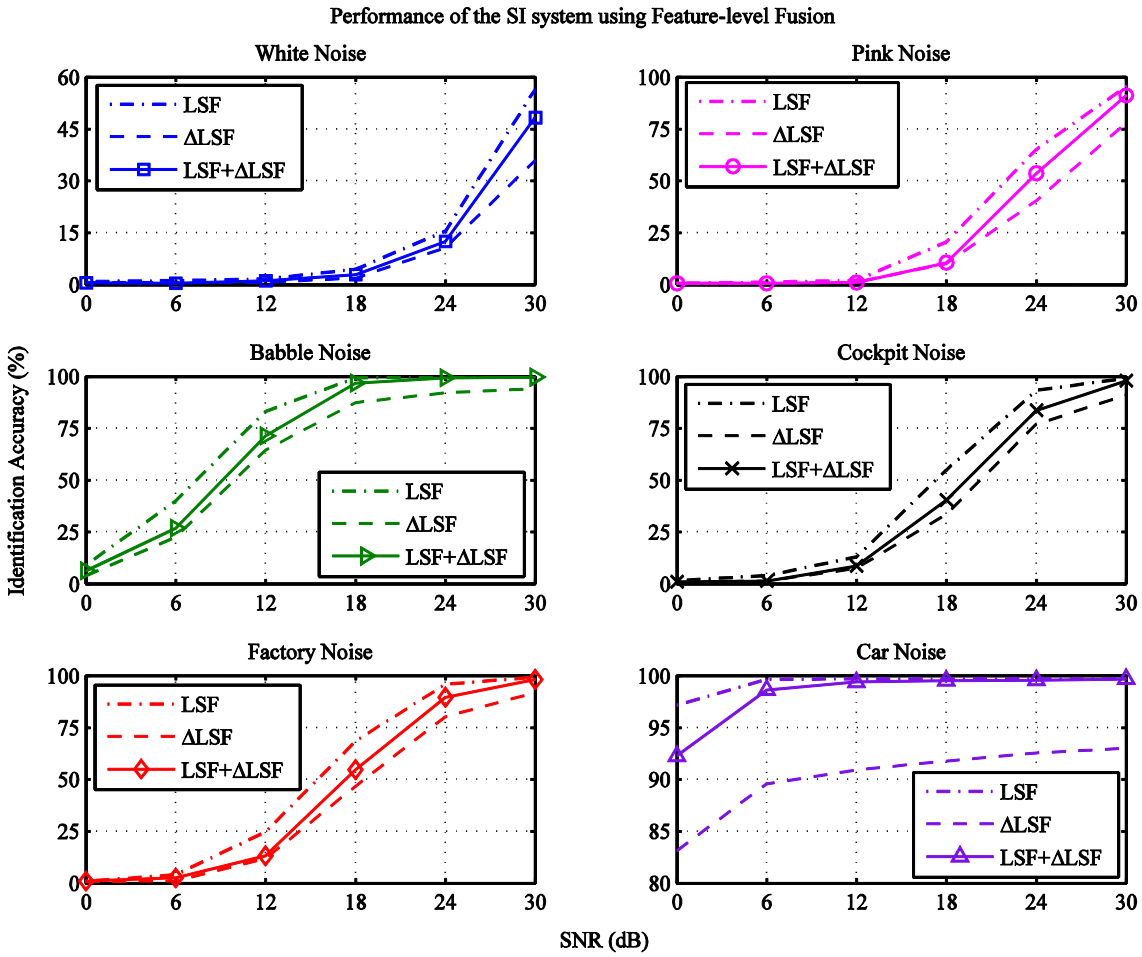

Fig. 6.10: Performance of the feature-level fusion based SI system in noise.

Under clean conditions, feature-level fusion provides the same performance as just using static LSF features. In addition, from Fig. 6.10, we observe that feature-level fusion degrades the performance in the presence of noise. Thus, it is clear that feature-level fusion is not beneficial, and its further pursuit is abandoned.

6.5.2 Score-level Fusion

In score-level fusion (SLF), each feature set is modeled separately, i.e. for each speaker, one GMM is created based on LSF features, and another GMM is based on ΔLSF features. During the evaluation phase, LSF and ΔLSF features are extracted from the test utterance and used to compute similarity scores using the corresponding GMMs of each speaker. These similarity scores are normalized to lie in the range [0, 1] using min-max normalization. Then, a weighted combination of the scores obtained from the LSF and ΔLSF GMMs is used to make the final decision.

$$\Phi_{s,f} = \alpha \Phi_{s,LSF} + (1-\alpha)\Phi_{s,\Delta LSF},\ 0 \le \Phi_{s,LSF}, \Phi_{s,\Delta LSF} \le 1, 0 < \alpha < 1 \qquad (6.4)$$

Then, the identified speaker using the score level fusion decision is:

$$\hat{s} = \arg\max_{1 \le s \le S} \Phi_{s,f} \tag{6.5}$$

In order to select the weight parameter α, a simple experiment is conducted. The value of α is varied from 0 to 1 and the identification accuracy is computed. The value of α that results in the highest identification accuracy is considered optimal. Under clean conditions, the results show that score-level fusion always performs lower than static LSF. Next, the experiment is repeated under white noise at 30 dB SNR. White noise is chosen to tune α since the SI system is worst affected by white noise. The hypothesis is that the value of α that is optimal for white noise will also work well under other noise types. The results are plotted in Fig. 6.11. Under noisy conditions, score-level fusion improves the performance of the SI system. Values of α in the range 0.6-0.8 are seen to result in the best performance under noisy conditions. Thus, α=0.7 is used in our score-level fusion system.

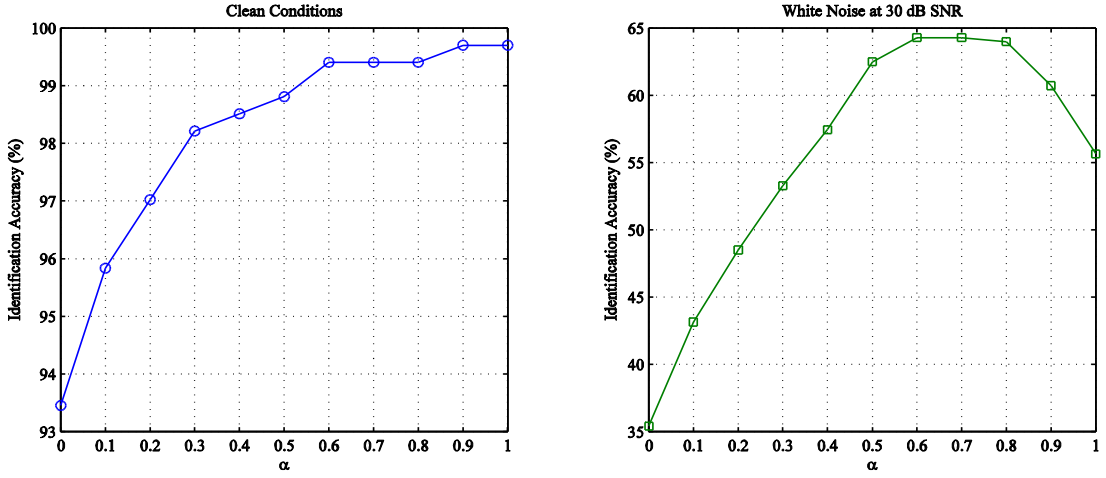

Fig. 6.11: Selection of the weight parameter for score-level fusion.

Next, the performance of the score-level fusion system is evaluated under clean as well as noisy conditions. The results are shown in Table 6.10 and in Table 6.11. Under clean conditions, we see that score-level fusion leads to a very slight deterioration in performance. This is because the ΔLSF based system performs much worse than the LSF based system under clean conditions.

Table 6.10: Performance of score-level fusion based SI system.

SNR (dB)	Identification Accuracy (%) of Score-level Fusion						
	White	Pink	Babble	Cockpit	Factory	Car	Average
30	65.00	96.61	99.58	98.81	99.17	99.40	93.10
24	17.74	71.37	99.29	95.00	96.79	99.40	79.93
18	3.27	19.40	98.99	60.12	73.15	99.46	59.07
12	1.13	1.43	88.75	13.39	26.07	99.70	38.41
6	0.83	1.07	47.02	2.44	3.51	99.70	25.76
0	0.60	0.60	12.08	1.13	0.77	99.35	19.09
Clean	99.40						

Table 6.11: Comparison of performance of score-level fusion based SI system.

SNR (dB)	Average Identification Accuracy (%)		
	LSF	ΔLSF	LSF+ ΔLSF (SLF)
30	91.63	80.53	93.10
24	78.18	65.49	79.93
18	57.86	45.18	59.07
12	37.31	29.43	38.41
6	24.94	19.31	25.76
0	18.29	14.91	19.09
Clean	99.70	93.45	99.40

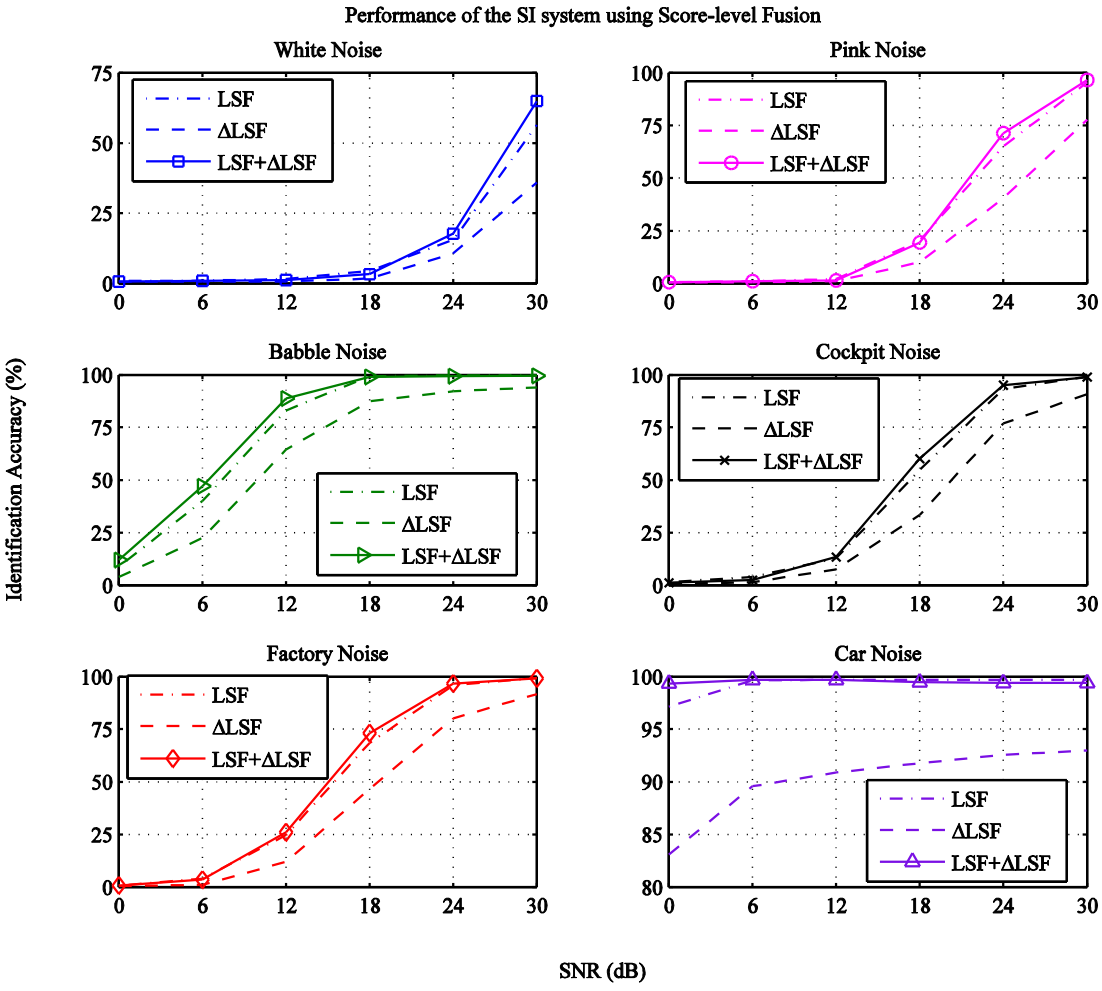

Fig. 6.12: Performance of the score-level fusion system under noise.

In noisy conditions, score-level fusion is seen to improve the identification accuracy. Under certain noise types, such as babble noise, the performance gain is much higher. Hence, the conclusion is that an SI system employing score-level fusion is more robust under noise.

6.5.3 Discriminative Power of Speech Zones

Now, for the score-level fusion based (LSF+ΔLSF) SI system, we investigate whether scoring exclusively on certain speech zones improves performance, under clean as well as noisy conditions. In particular, we are interested in measuring the speaker-discriminative power of transitions into and out of vowels.

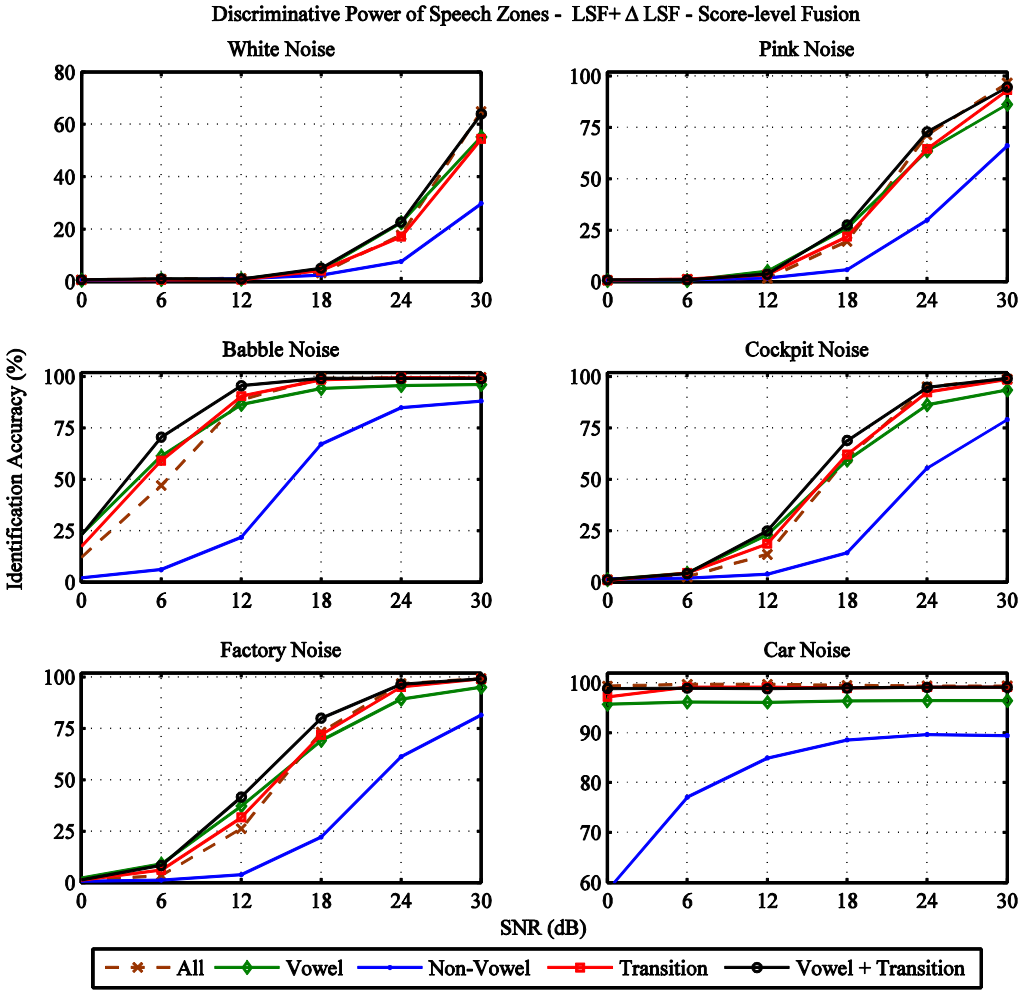

Fig. 6.13: Discriminative power of speech zones for a score-level fusion based system.

Again, we infer that in clean and high SNR conditions, frame-level selection is not advantageous. It is more effective to utilize all available frames. However, frame-level selection is very beneficial under lower SNR conditions. Scoring only on certain speech zones boosts the identification accuracy substantially under high noise levels.

Among the different speech zones, non-vowel frames appear to be the least speaker-discriminative in clean as well as noisy conditions. Under high SNR conditions, transition frames contain the most speaker-specific information. Scoring only on transition frames results in nearly the same performance as scoring on all frames. As SNR decreases, the curves of the transition and vowel

frames are seen to cross over. At low SNR, scoring only on vowel frames produced the best results among the individual categories. This tells us that at lower SNR conditions, vowel frames are the most-speaker discriminative. However, the highest identification accuracy was almost always obtained by scoring on a combination of vowel and transition frames. In particular, under low SNR conditions, scoring on a combination of transition and vowel frames resulted in a tremendous improvement in performance.

As observed earlier, the SI system is worst affected by white noise. At 30 dB SNR, frame-level selection did not result in any performance improvement. However, at 24 dB and 18 dB SNR, scoring on a combination of transition and vowel frames outperforms scoring on all frames by a considerable margin. From 12 dB and below, we are not able to observe any clear differences between the different categories as the system is severely affected by noise.

A similar result is observed under pink noise. At 30 dB SNR, frame-level selection does not lead to any performance improvement. At SNRs in the range 12 dB-24 dB, frame-level selection begins to result in major performance gains. The best performance was obtained by scoring on a combination of vowel and transition frames. From below 12 dB, the performance of the system is severely affected and we do not observe any clear trends.

In the presence of babble noise, down to 18 dB SNR, the best performance is obtained by scoring on all speech frames. From 12 dB SNR and below, scoring on transition and vowel frames is observed to be the best approach, resulting in marked performance improvements over scoring on all frames.

Similarly, under cockpit and factory noise, at SNR levels of 30 dB - 24 dB, the best performance is obtained by scoring on all frames. From 18 dB SNR and below, frame-level selection boosts the identification accuracy. Under cockpit noise as well as factory noise, by scoring selectively on vowel and transition frames, significant improvements are obtained at SNRs of 18 dB to 6 dB.

Under car noise, a different behavior is observed since the SI system is extremely robust against this noise type. In this case, the transition frames are the most speaker-discriminative, and do not lose their discriminative power under noise. Scoring on transition frames only results in nearly the same performance as the baseline, all the way down to 0 dB SNR. Frame-level selection does not result in any performance gain.

Thus, we conclude that the score-level fusion system, which combines information from the static LSF and delta-LSF classifiers, is able to extract complementary information from transition regions, as well as steady vowel regions, resulting in more robust decisions. In addition, we are able to improve the performance of the score-level fusion system further by using only vowel and transition zones of speech for scoring, during the identification phase.

Table 6.12: Discriminative power of speech zones for the score-level fusion based SI system.

Noise Type	SNR (dB)	Identification Accuracy (%) of the score-level fusion based SI system				
		All	Non-Vowel	Transition	Vowel	Vowel + Transition
Clean		99.40	89.58	99.11	96.43	99.11
White	30	65.00	29.82	54.46	55.18	64.05
	24	17.74	7.68	17.08	**22.44**	**22.74**
	18	3.27	2.38	**4.17**	**4.52**	**5.06**
	12	1.13	1.01	0.65	0.89	0.95
	6	0.83	0.65	0.65	0.77	**1.07**
	0	0.60	0.54	0.60	0.54	0.60
Pink	30	96.61	66.13	93.10	86.25	94.70
	24	71.37	29.88	64.58	63.57	**72.92**
	18	19.40	5.77	**21.85**	**26.07**	**27.50**
	12	1.43	1.73	**3.39**	**4.94**	**3.63**
	6	1.07	0.83	1.07	0.48	0.83
	0	0.60	0.48	0.60	0.48	**0.83**
Babble	30	99.58	88.10	99.35	96.19	99.11
	24	99.29	84.82	99.40	95.65	99.11
	18	98.99	67.08	98.39	94.23	98.99
	12	88.75	21.67	**90.54**	86.49	**95.48**
	6	47.02	5.95	**58.99**	**61.49**	**70.54**
	0	12.08	2.08	**17.32**	**23.10**	**23.10**
Cockpit	30	98.81	78.99	98.57	93.51	99.11
	24	95.00	55.36	92.32	86.31	94.82
	18	60.12	14.11	**62.02**	59.29	**68.81**
	12	13.39	3.81	**18.51**	**22.98**	**24.82**
	6	2.44	1.79	**4.29**	**4.29**	**4.11**
	0	1.13	**1.43**	0.83	1.13	1.13
Factory	30	99.17	81.43	99.11	95.06	98.99
	24	96.79	61.19	95.12	89.23	96.55
	18	73.15	22.08	71.85	69.17	**79.94**
	12	26.07	3.69	**31.61**	**37.32**	**41.61**
	6	3.51	1.01	**6.13**	**8.99**	**8.21**
	0	0.77	0.54	**1.13**	**2.02**	**1.31**
Car	30	99.40	89.46	99.11	96.43	99.11
	24	99.40	89.64	99.11	96.43	99.11
	18	99.46	88.57	98.99	96.37	98.99
	12	99.70	84.94	99.11	96.13	98.81
	6	99.70	77.08	99.11	96.19	98.93
	0	99.35	58.39	97.20	95.71	98.87

6.6 CONCLUSIONS

From the experiments conducted above, we arrive at the following inferences:

1. Frame selection during scoring is not advantageous under high SNR conditions. Under low SNR conditions, frame selection improves identification accuracy.
2. For a static LSF based SI system, transition regions are most speaker discriminative in high SNR conditions. Under low SNR conditions, vowel regions are most speaker discriminative. When only vowel frames are selected for scoring, a significant performance improvement over the baseline is observed.
3. An SI system based on delta LSF features operates in a complementary manner. However, the identification accuracy is worse than that of a static LSF based system.
4. Feature-level fusion of LSF and delta-LSF features is not beneficial as this causes a degradation in performance under noisy conditions.
5. Score-level fusion of LSF and delta-LSF based classifiers improves the robustness of the SI system under noise.
6. For the above score-level fusion based SI system, transition regions are most speaker discriminative under high SNR conditions. Under low SNR conditions, vowel regions are most speaker discriminative. In noisy conditions, scoring exclusively on a combination of both transition and vowel frames results in significant performance gains.

Thus, we propose a speaker identification system which fuses information from static LSF and dynamic ΔLSF based classifiers at the score-level, and utilizes a combination of steady vowel and CV/VC transition zones of speech for scoring during the identification phase.

The relative improvement in identification accuracy obtained by the proposed SI system over the baseline is presented in Table 6.13 and Table 6.14. The proposed system is denoted by LSF+ ΔLSF (Vowel+Transition), and the baseline LSF based SI system which does not perform any frame-level selection is denoted by LSF (All). Under clean conditions, the proposed SI system performs nearly as well as the baseline system. Only a very slight performance degradation is observed. Under noisy conditions, significant performance gains are obtained.

Table 6.13: Comparison of average identification accuracy over various noise types obtained by the proposed system.

SNR (dB)	**Average Identification Accuracy (%)**		**Relative Performance Improvement (%)**
	LSF (All)	**LSF+ΔLSF (SLF) (Vowel+Transition)**	
30	91.63	92.51	0.96
24	78.18	80.87	3.44
18	57.86	63.21	9.25
12	37.31	44.22	18.52
6	24.94	30.62	22.77
0	18.29	20.97	14.65
Clean	99.70	99.11	-0.59

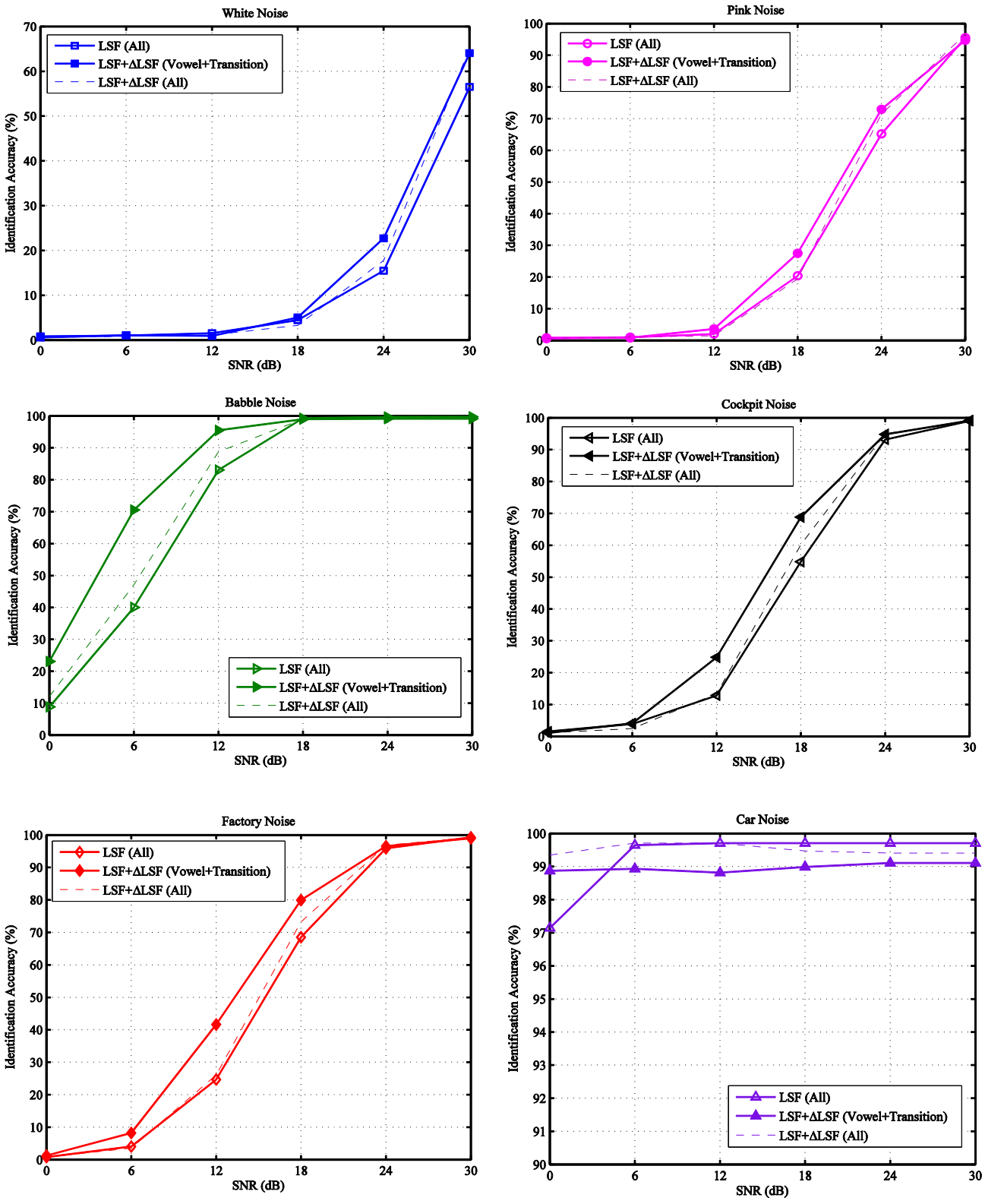

Fig. 6.14: Performance improvement over the baseline system by using score-level fusion and scoring exclusively on vowel and transition frames.

Table 6.14: Performance improvement by using the proposed SI system over the baseline under different noise conditions.

Noise Type	SNR (dB)	Identification Accuracy (%)		
		LSF (All)	**LSF+ ΔLSF (All)**	**LSF+ ΔLSF (Vowel + Transition)**
Clean		**99.70**	99.40	99.11
White	30	56.49	**65.00**	64.05
	24	15.48	17.74	**22.74**
	18	4.46	3.27	**5.06**
	12	**1.55**	1.13	0.95
	6	1.01	0.83	**1.07**
	0	**0.83**	0.60	0.60
Pink	30	95.54	**96.61**	94.70
	24	65.18	71.37	**72.92**
	18	20.36	19.40	**27.50**
	12	2.02	1.43	**3.63**
	6	1.01	**1.07**	0.83
	0	0.60	0.60	**0.83**
Babble	30	**99.70**	99.58	99.11
	24	**99.64**	99.29	99.11
	18	**99.29**	98.99	98.99
	12	83.04	88.75	**95.48**
	6	40.00	47.02	**70.54**
	0	8.87	12.08	**23.10**
Cockpit	30	**99.05**	98.81	99.11
	24	93.21	**95.00**	94.82
	18	54.82	60.12	**68.81**
	12	12.86	13.39	**24.82**
	6	3.93	2.44	**4.11**
	0	**1.55**	1.13	1.13
Factory	30	**99.29**	99.17	98.99
	24	95.89	**96.79**	96.55
	18	68.51	73.15	**79.94**
	12	24.70	26.07	**41.61**
	6	4.05	3.51	**8.21**
	0	0.77	0.77	**1.31**
Car	30	**99.70**	99.40	99.11
	24	**99.70**	99.40	99.11
	18	**99.70**	99.46	98.99
	12	**99.70**	99.70	98.81
	6	99.64	**99.70**	98.93
	0	97.14	**99.35**	98.87

From Fig. 6.14, we see that under babble noise, very large performance improvements are observed for SNRs in the range 0-12 dB. Under babble noise at 12 dB, the identification accuracy increases from 83.04% to 95.48% using the proposed SI system. This corresponds to a relative improvement of 14.98%. Similarly, at 6 dB SNR, the performance increases from 40% to 70.54%, i.e. a 76.35% relative improvement. At 0 dB SNR, the identification accuracy is boosted from 8.87% to 14.43%, which is almost a two-fold improvement.

In the presence of factory noise, the proposed SI system outperforms the baseline considerably at lower SNRs. At 18 dB, the proposed system results in 79.94% accuracy as opposed to 68.51% for the baseline. This corresponds to a relative improvement of 16.68%. Similarly at 12 dB, the proposed system has an accuracy of 41.61% - which is a relative improvement of 68.46% compared to the baseline. At 6 dB, the identification accuracy increases two-fold from 4.05% to 8.21%.

Under cockpit noise, at 18 dB the proposed system results in an identification accuracy of 68.81%, whereas the baseline performance is at 54.82%. This corresponds to a relative performance improvement of 25.52%. Similarly at 12 dB SNR, the performance improves from 12.86% to 24.82%, i.e. a 93% relative improvement. At 6 dB SNR, a relative improvement of 4.58% is achieved.

Similarly, for speech degraded by white noise, relative performance improvements of 13.38%, 46.9%, and 13.45% are obtained at SNRs of 30 dB, 24 dB, and 18 dB respectively. Under pink noise at 24 dB and 18 dB, relative performance improvements of 11.87% and 35.07% respectively are obtained. Since the LSF based SI system is already very robust against car noise, no major further performance improvement is observed in this case.

Thus, by employing a score-level fusion of LSF and delta-LSF based classifiers, and selecting only vowel and CV/VC transition frames for similarity scoring, we are able to improve the robustness of an LSF based SI system under various noise conditions.

7 SPEAKER VERIFICATION EXPERIMENTS

In this chapter, the set-up and performance evaluation of our baseline speaker verification system using LSF features is explained. The demonstration in Section 6.5.2 showed that score-level fusion is more effective than feature-level fusion; this investigation is therefore limited to score-level fusion of static and dynamic LSF features for speaker verification. Also, none of the experiments that were conducted to investigate the relative importance of vowel, non-vowel, and transition speech zones is repeated for the purpose of speaker verification. The working hypothesis is that the speaker-discriminative power of different regions will remain the same, irrespective of whether the application is speaker identification or speaker verification. Once the score-level fusion based SV system is set up, the performance improvement obtained by scoring exclusively on a combination of vowel and transition frames under noisy conditions is evaluated.

7.1 EXPERIMENTAL SETUP

The experimental setup of our speaker verification system is outlined below.

7.1.1 Speaker Set

The set of 168 speakers in the TEST directory of the TIMIT corpus is selected for the training and evaluation of our SV system. This set consists of 112 male speakers and 56 female speakers. The training set for each speaker consists of the 5 SX and the 3 SI sentences. All the sentences in the training set are normalized and concatenated to produce approximately 24 seconds of training speech. The test set consists of the two SA sentences.

7.1.2 Universal Background Model

The UBM is trained using the set of 462 speakers present in the TRAIN directory of the TIMIT corpus. All sentences from all the speakers are pooled for UBM training. The training speech is divided in 20 msec long frames, with a frame shift of 10 msec. LSF feature vectors of order p=20 are extracted from the speech frames. A 256 component GMM is trained using these features, as explained in Chapter 4. Nodal, diagonal covariance matrices are used.

7.1.3 Speaker Enrollment

For each of the 168 speakers in the TEST directory, an adapted Gaussian Mixture Model is created using the following procedure. The training speech is divided in 20 msec long frames, with a frame shift of 10 msec. LSF feature vectors of order p=20 are extracted from the speech frames. For each speaker, these feature vectors are used to create a maximum a posteriori (MAP) adapted GMM from the UBM. Only the UBM means are adapted with a relevance factor of r=16, as outlined in Section 4.3.3.

7.1.4 Performance Evaluation

7.1.4.1 Clean Conditions

The performance of the LSF based speaker verification system is evaluated under clean conditions as shown in Table 7.1. For each of the 168 speakers, the two SA sentences in the test set are used in individual target trials, of approximately 3 seconds each. Thus, a total of $168 \times 2 = 336$ target trials is conducted. Each test utterance is divided in 20 msec long frames, with a frame shift of 10 msec. LSF feature vectors of dimension p=20 are extracted from the speech frames and used for evaluation. These features are used to obtain the log-likelihood ratio by scoring across the target GMM and also the UBM. The scores of all the target trials are collected and stored.

Next, the 2 SA sentences in the test set of each speaker are used as impostor trials with the rest of the speaker models. Thus, $168 \times 2 \times 167 = 56112$ impostor trials are conducted. The scores of all the impostor trials are collected and stored in another pool. Next, a decision threshold θ is swept over the two sets of scores and the probability of miss and probability of false alarm are computed for each threshold. The Equal Error Rate (EER), i.e. the point at which the probabilities of miss and false alarm are equal is used as a performance measure. Detection Error Tradeoff (DET) curves, explained in Section 4.3.5, are also plotted.

Table 7.1: Target and impostor trials

# Speakers	# Target Trials	# Impostor Trials	# Total Trials
168	168 x 2 = 336	168 x 167 x 2 = 56112	56112 + 336 = 56448

7.1.4.2 Noisy Conditions

Next, the performance of the LSF based SV system is evaluated in the presence of different kinds of stationary/non-stationary background noise. The noise signals were obtained from the Signal Processing Information Base (SPIB) noise dataset [95]. The performance evaluation is conducted for speech contaminated by white, pink, babble, cockpit, factory, and car noise, at SNRs from 30 dB to 0 dB, in steps of 6 dB.

For a particular noise type at a particular SNR, the evaluation is conducted in the following manner. For every speaker, each of the two SA test sentences is corrupted with a random realization of noise. The noisy speech signals were created by selecting a random segment from the noise file, scaling the noise to obtain the appropriate SNR, and then adding it to the clean speech signal. Then, using these noisy test segments, a total of 336 target trials and 56112 impostor trials is conducted, in the same manner as described above. The Equal Error Rate (EER), as described in Section 4.3.5, is used as a performance measure.

7.2 SPEAKER VERIFICATION USING LSF FEATURES

The parameters of our baseline speaker verification system are outlined in the following table.

Table 7.2: Parameters of the LSF based SV system.

Parameter	Description
Number of speakers (*S*)	168
Training set of each speaker	All SX, SI sentences (~3 seconds x 8)
Test set of each speaker	SA sentences (~3 seconds x 2)
Feature Type	LSF
Feature Dimension/Order (*p*)	20
Frame Length (*L*)	20 msec
Frame Shift (δ)	10 msec
Number of GMM Components (*M*)	256 (UBM adapted GMM)
GMM Covariance Type	Nodal and Diagonal

7.2.1 Performance Evaluation

In this section, the performance evaluation of the LSF based speaker verification system is presented. Under clean conditions, the LSF based SV system has an Equal Error Rate (EER) of 0.89%. The DET curve of the SV system is shown in Fig. 7.1.

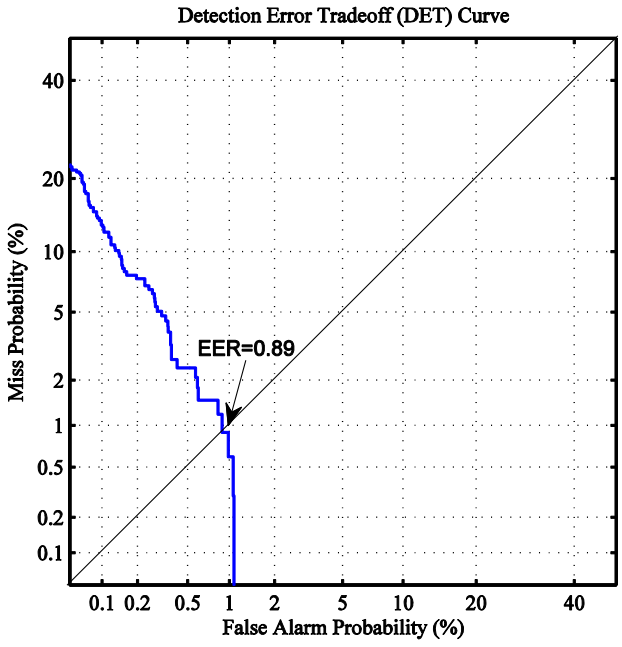

Fig. 7.1: DET Curve of the LSF based SV system.

Next, the performance of the LSF based speaker verification system is evaluated in the presence of different types of background noise, as described in Section 7.1.4.2. The results are presented in Table 7.3, and also represented in Fig. 7.2.

Table 7.3: Performance of the LSF based SV system under noisy conditions.

SNR (dB)	Equal Error Rate (%)						
	White	**Pink**	**Babble**	**Cockpit**	**Factory**	**Car**	**Average**
30	6.55	2.11	0.94	1.19	1.19	0.89	2.15
24	10.71	6.25	1.01	2.81	1.81	0.90	3.91
18	21.13	10.89	1.29	9.23	4.85	0.99	8.06
12	36.61	23.28	2.70	21.51	13.81	0.98	16.82
6	45.54	37.50	9.13	37.10	29.18	1.15	26.60
0	48.51	45.83	23.21	44.94	40.18	1.60	34.05
Clean	0.89						

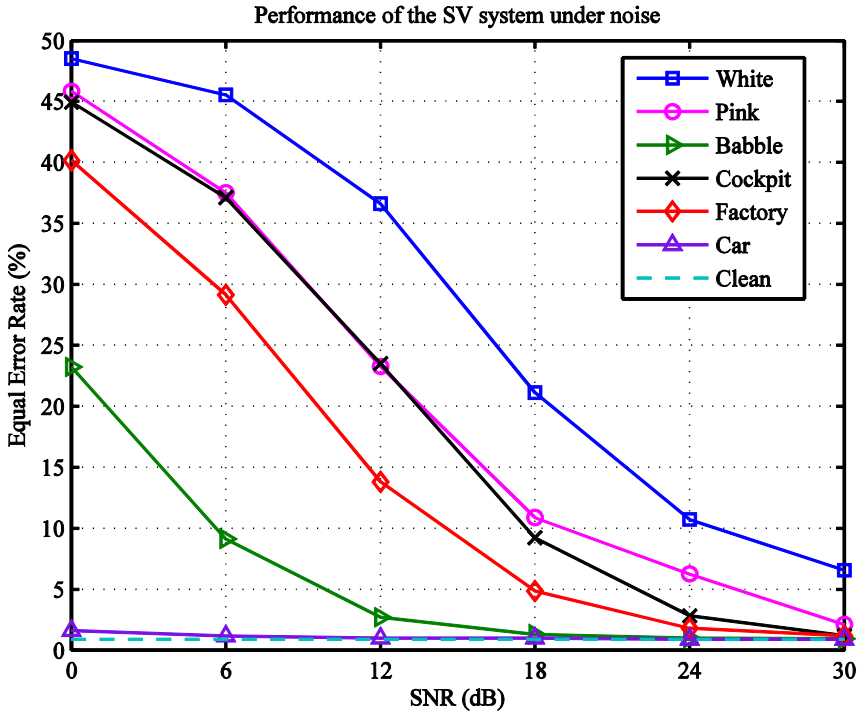

Fig. 7.2: Performance of the LSF based SV system under noise.

From Fig. 7.2, we observe that the performance of the LSF based SV system degrades considerably in the presence of most noise types. Like the SI system in Chapter 6, the SV system also seems to be most immune to car noise, for which the EER begins to increase only at 0 dB SNR. The SI system is also reasonably robust against babble noise. A major increase in EER is observed starting at SNR levels below 18 dB. The EER is worst affected by white noise. At 0 dB SNR, the EER is observed to reach almost 50%, which is equivalent to making a random guess.

Thus, we conclude that while the LSF based SV system provides near perfect performance under clean conditions, the performance degrades considerably under most types of noise. Next, the performance of an SV system that used dynamic LSF features is investigated.

7.3 SPEAKER VERIFICATION USING DELTA-LSF FEATURES

Our baseline speaker verification system in the previous section utilized 'static' LSF feature vectors. In order to incorporate more dynamic/transitional information, first-order delta-LSF (ΔLSF) features are extracted using the same method as described in Chapter 6.

7.3.1 Performance Evaluation

In this section, the effect on the verification performance of using ΔLSF features instead of static LSF features is analyzed. In order to do this, a speaker verification system is set up and evaluated using the same parameters as in Table 7.2Table 6.1. The only difference is the use of ΔLSF features of order p=20, instead of LSF features, for creating the UBM, adapting the speaker models, and subsequent performance evaluation. A regression window parameter K=2 is chosen for computing the ΔLSF features. The results of the evaluation of SV performance based on ΔLSF features in clean as well as noisy conditions are presented in Table 7.4. A comparison of the DET curves of the LSF and ΔLSF based systems in clean conditions is shown in Fig. 6.8.

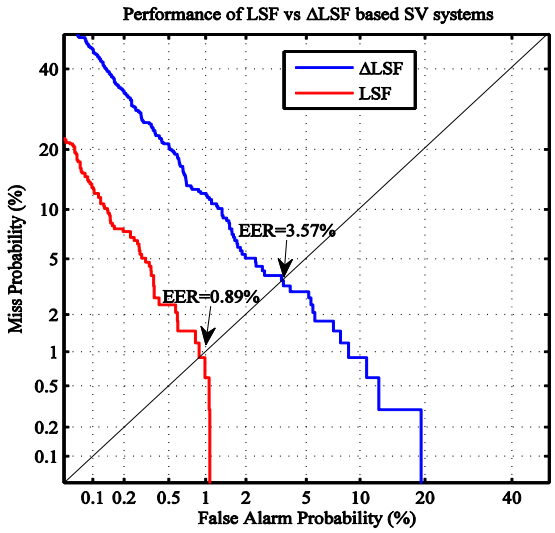

Fig. 7.3: DET curves of LSF vs delta-LSF based SV systems in clean conditions.

Table 7.4: Performance of the delta-LSF based SV system in noisy conditions.

SNR (dB)	Equal Error Rate (%) of the ΔLSF based SV system						
	White	Pink	Babble	Cockpit	Factory	Car	Average
30	8.10	4.46	3.03	3.74	3.72	3.43	4.41
24	15.58	7.44	3.66	5.07	5.06	3.27	6.68
18	25.01	17.21	4.00	11.61	8.79	3.41	11.67
12	36.61	28.54	6.55	24.39	17.86	3.57	19.59
6	46.05	39.88	12.50	36.02	29.48	3.57	27.92
0	48.81	45.54	24.70	43.62	41.67	3.57	34.65
Clean	3.57						

Under background noise, the behavior of the ΔLSF based SV system is similar to that of the LSF based SV system. The performance seems to be most immune to car noise, followed by babble noise. The Equal Error Rate is worst affected by pink and white noise.

A comparison between the average performance of LSF and ΔLSF based SV systems under noise is shown in Table 7.5 and also in Fig. 7.4. The performance obtained by using ΔLSF features is observed to be consistently worse than that using LSF features, under clean as well as noisy conditions.

Table 7.5: Performance comparison of LSF vs delta-LSF based SV systems.

SNR (dB)	Average Equal Error Rate (%)	
	LSF	ΔLSF
30	2.15	4.41
24	3.91	6.68
18	8.06	11.67
12	16.82	19.59
6	26.60	27.92
0	34.05	34.65
Clean	0.89	3.57

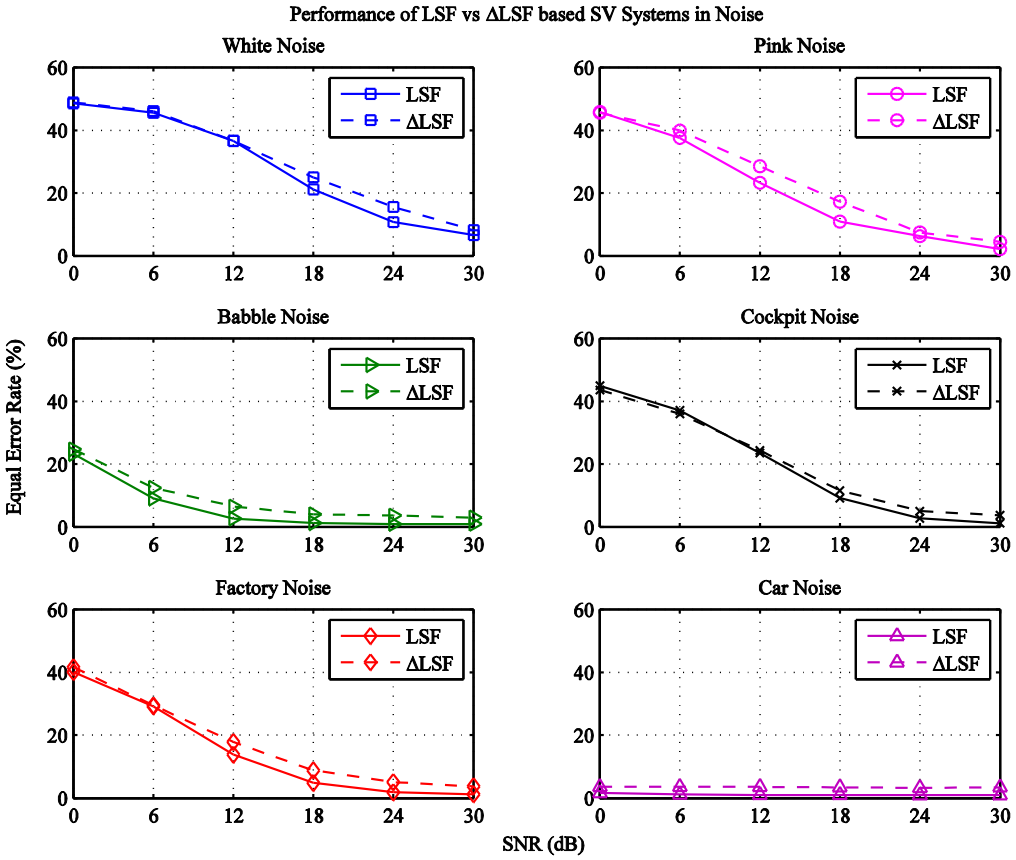

Fig. 7.4: Performance of LSF vs delta-LSF based SV systems in noise.

Thus, we conclude that ΔLSF features are not ideal for speaker verification when used on their own. In the next section, the score-level fusion of LSF and ΔLSF classifiers is explored for its potential to make a more robust speaker verification system.

7.4 FUSION OF INFORMATION FROM LSF AND DELTA-LSF FEATURES

In Chapter 6, for speaker identification, we explored two commonly used fusion techniques, namely feature-level fusion and score-level fusion. It was found that score-level fusion was more effective than feature-level fusion. Hence, a score-level fusion technique is selected for speaker verification as well.

7.4.1 Score-level Fusion

In score-level fusion (SLF), each feature set is modeled separately. First, using the set of 462 speakers from the TRAIN directory of the TIMIT corpus, one UBM is trained based on LSF features, and another UBM is trained based on ΔLSF features. For each speaker, one adapted GMM is created from the LSF based UBM, and another adapted GMM is created from the ΔLSF feature based UBM. During the evaluation phase, LSF and ΔLSF features are extracted from the test utterance and used to compute log-likelihood ratios using the corresponding GMMs of each speaker and the corresponding UBMs, as shown in equations (4.43) and (4.44). Then,

$$\Lambda_{LSF}(X) = \Phi_{s,LSF} - \Phi_{ubm,LSF} \tag{7.1}$$

$$\Lambda_{\Delta LSF}(X) = \Phi_{s,\Delta LSF} - \Phi_{ubm,\Delta LSF} \tag{7.2}$$

These log-likelihood ratios are normalized to lie in the range [0, 1]. Then, a weighted combination of the log-likelihood ratios obtained from LSF and ΔLSF GMMs is used to make the final decision.

$$\begin{aligned} \Lambda_f(X) &= \alpha\Lambda_{LSF}(X) + (1-\alpha)\Lambda_{\Delta LSF}(X) \\ 0 &\le \Lambda_{LSF}(X), \Lambda_{\Delta LSF}(X) \le 1 \\ 0 &< \alpha < 1 \end{aligned} \tag{7.3}$$

Based on our speaker identification experiments in Section 6.5.2, the weight parameter α is chosen to be 0.7 in our score-level fusion system.

Next, the performance of the score-level fusion system is evaluated under clean as well as noisy conditions. The results are shown in Table 7.6 and in Table 7.7. The DET curves of the LSF, ΔLSF and score-level fusion based systems under clean and noisy conditions are shown in Fig. 7.5. From the figure, we observe that the score-level fusion based system performs very similar to the static LSF based system in clean conditions. No clear performance improvement is observed. However, under noisy conditions, we see that the score-level fusion system consistently performs better than the LSF as well as the ΔLSF based systems.

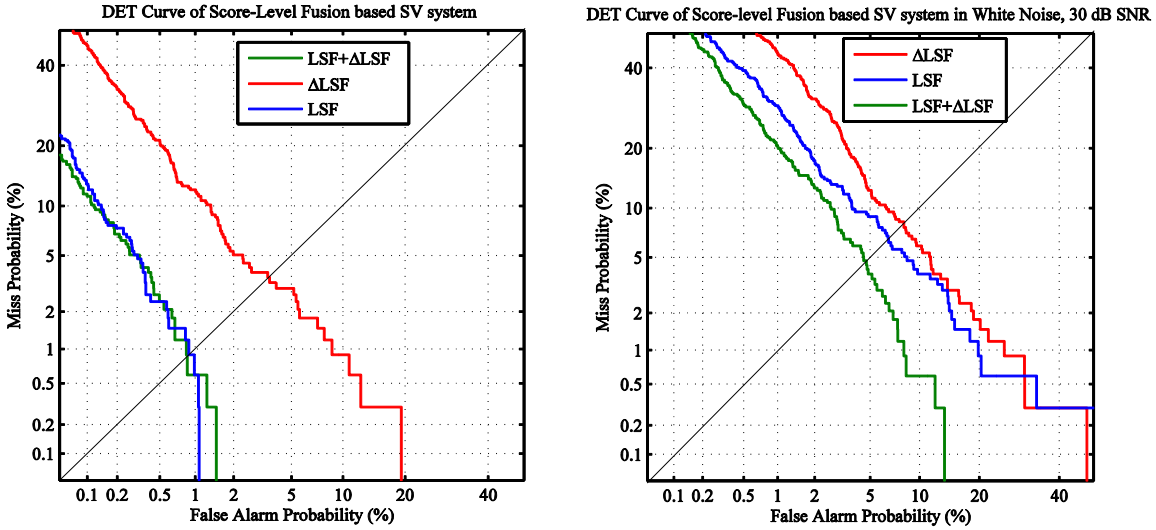

Fig. 7.5: DET curve of score-level fusion based SV system in clean vs noisy conditions conditions.

Table 7.6: Performance of score-level fusion based SI system.

SNR (dB)	Equal Error Rate (%) of Score-level Fusion						
	White	Pink	Babble	Cockpit	Factory	Car	Average
30	4.71	1.59	0.86	1.06	1.01	0.87	1.68
24	8.63	4.04	0.84	1.91	1.58	0.89	2.98
18	17.33	8.92	0.87	6.91	4.21	0.89	6.52
12	33.93	19.35	1.79	16.96	10.41	0.89	13.89
6	43.75	35.40	6.55	31.25	24.11	0.87	23.66
0	46.54	43.23	19.05	43.61	38.69	1.19	32.05
Clean	0.86						

Table 7.7: Comparison of performance of score-level fusion based SI system.

SNR (dB)	Average Equal Error Rate (%)		
	LSF	ΔLSF	LSF+ ΔLSF (SLF)
30	2.15	4.41	1.68
24	3.91	6.68	2.98
18	8.06	11.67	6.52
12	16.82	19.59	13.89
6	26.60	27.92	23.66
0	34.05	34.65	32.05
Clean	0.89	3.57	0.86

In Fig 7.6, the EER of the LSF, ΔLSF, and score-level fusion based systems are compared under different noise types at different SNR levels.

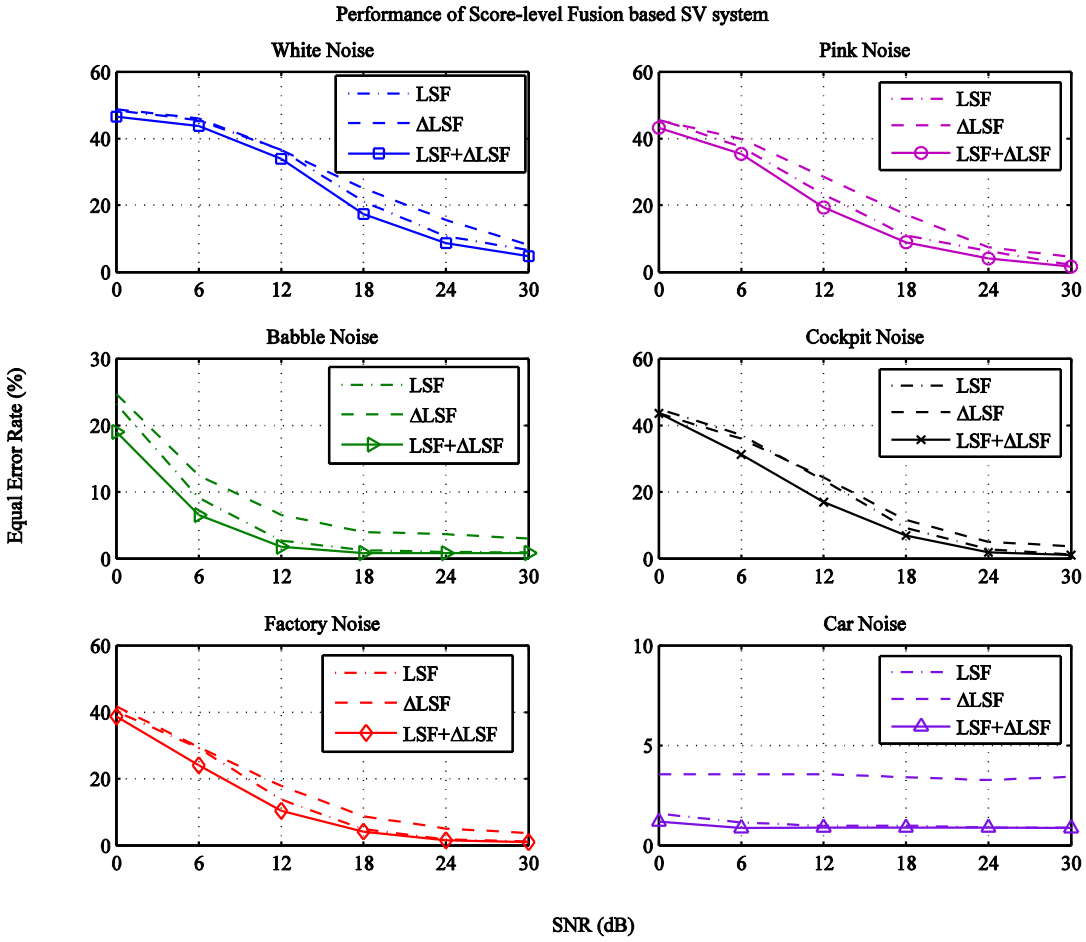

Fig. 7.6: Performance of the score-level fusion based SV system.

We observe that EER of the score-level fusion based system is consistently lower under all noise conditions. Hence, we conclude that an SV system employing score-level fusion performs comparably to the static LSF based SV system under clean conditions, and is more robust against noise than the static LSF based SV system.

7.4.2 Discriminative Power of Vowels and Transitions

For the score-level fusion based (LSF+ΔLSF) SV system, we investigate whether scoring exclusively on vowel and transition frames improves performance, under clean as well as noisy conditions. During the verification phase, the speech frames are classified into three categories – vowel, non-vowel, and transition, using the same methodology as proposed in Chapter 6.

In the speaker identification experiments, we proposed a robust SI system, which combines information from the static LSF and delta-LSF classifiers at the score level, and also utilizes only a combination of transition and vowel regions for scoring. Now, the same concept is tested in speaker verification. For the score-level fusion based SV system, we study the performance improvements obtained by utilizing only vowel and transition frames for scoring.

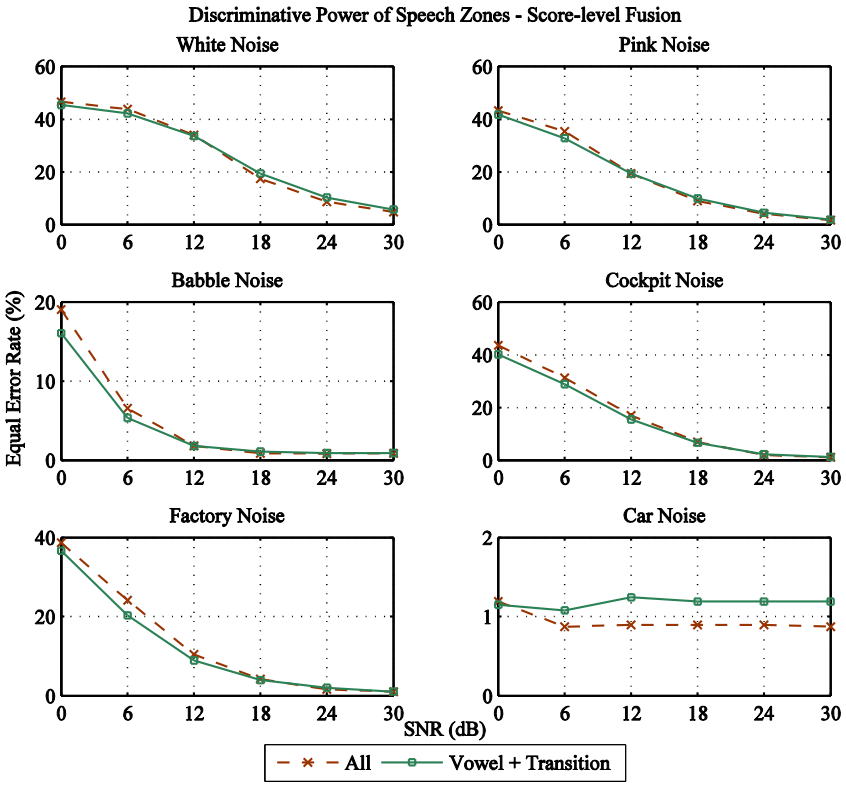

Fig. 7.7: Discriminative power of speech zones for the score-level fusion based SV system.

From Fig. 7.7, we see that scoring on a combination of vowel and transition frames only increases the EER of the score-level fusion based SV system at SNRs above 18 dB. Thus, for the score-level fusion based SV system, frame-level selection does not result in any benefit in high SNR conditions.

Under white noise, scoring on transition and vowel frames only improves SV performance from 12 dB SNR and below. Similarly, under pink noise, scoring on transition and vowel frames only resulted in a decrease in EER from 6 dB and below. In the presence of babble noise as well, scoring on transition and vowel frames only resulted in a decrease in EER from 6 dB and below. Under cockpit and factory noise, scoring on transition and vowel frames only resulted in a decrease in EER from 18 dB and below.

Thus, by selecting only vowel and CV/VC transition frames for similarity scoring, we are able to improve the robustness of the score-level fusion based SV system only under very low SNR conditions. Recall that, for the score-level fusion based SI system in Chapter 6, we were able to obtain better performance by selecting vowel and transition frames for scoring, even under high SNR conditions. In addition, the relative performance improvements are much higher in the case of speaker identification than in the case of speaker verification.

Table 7.8: Discriminative power of vowel and transition frames for the score-fusion based SV system.

Noise Type	SNR (dB)	Equal Error Rate (%)		
		LSF (All)	**LSF+ ΔLSF (All)**	**LSF+ ΔLSF (Vowel+Transition)**
Clean		0.89	0.86	1.19
White	30	6.55	**4.71**	5.65
	24	10.71	**8.63**	10.21
	18	21.13	**17.33**	19.40
	12	36.61	33.93	**33.63**
	6	45.54	43.75	**42.15**
	0	48.51	46.54	**45.34**
Pink	30	2.11	**1.59**	1.85
	24	6.25	**4.04**	4.46
	18	10.89	**8.92**	9.86
	12	23.28	**19.35**	**19.35**
	6	37.50	35.40	**32.74**
	0	45.83	43.23	**41.74**
Babble	30	0.94	**0.86**	0.89
	24	1.01	**0.84**	0.88
	18	1.29	**0.87**	1.07
	12	2.70	**1.79**	**1.79**
	6	9.13	6.55	**5.36**
	0	23.21	19.05	**16.04**
Cockpit	30	1.19	**1.06**	1.19
	24	2.81	**1.91**	2.16
	18	9.23	6.91	**6.55**
	12	21.51	16.96	**15.48**
	6	37.10	31.25	**28.80**
	0	44.94	43.61	**40.18**
Factory	30	1.19	**1.01**	1.03
	24	1.81	**1.58**	1.99
	18	4.85	4.21	**3.92**
	12	13.81	10.41	**8.93**
	6	29.18	24.11	**20.30**
	0	40.18	38.69	**36.61**
Car	30	0.89	**0.87**	1.19
	24	0.90	**0.89**	1.19
	18	0.99	**0.89**	1.19
	12	0.98	**0.89**	1.24
	6	1.15	**0.87**	1.08
	0	1.60	1.19	**1.15**

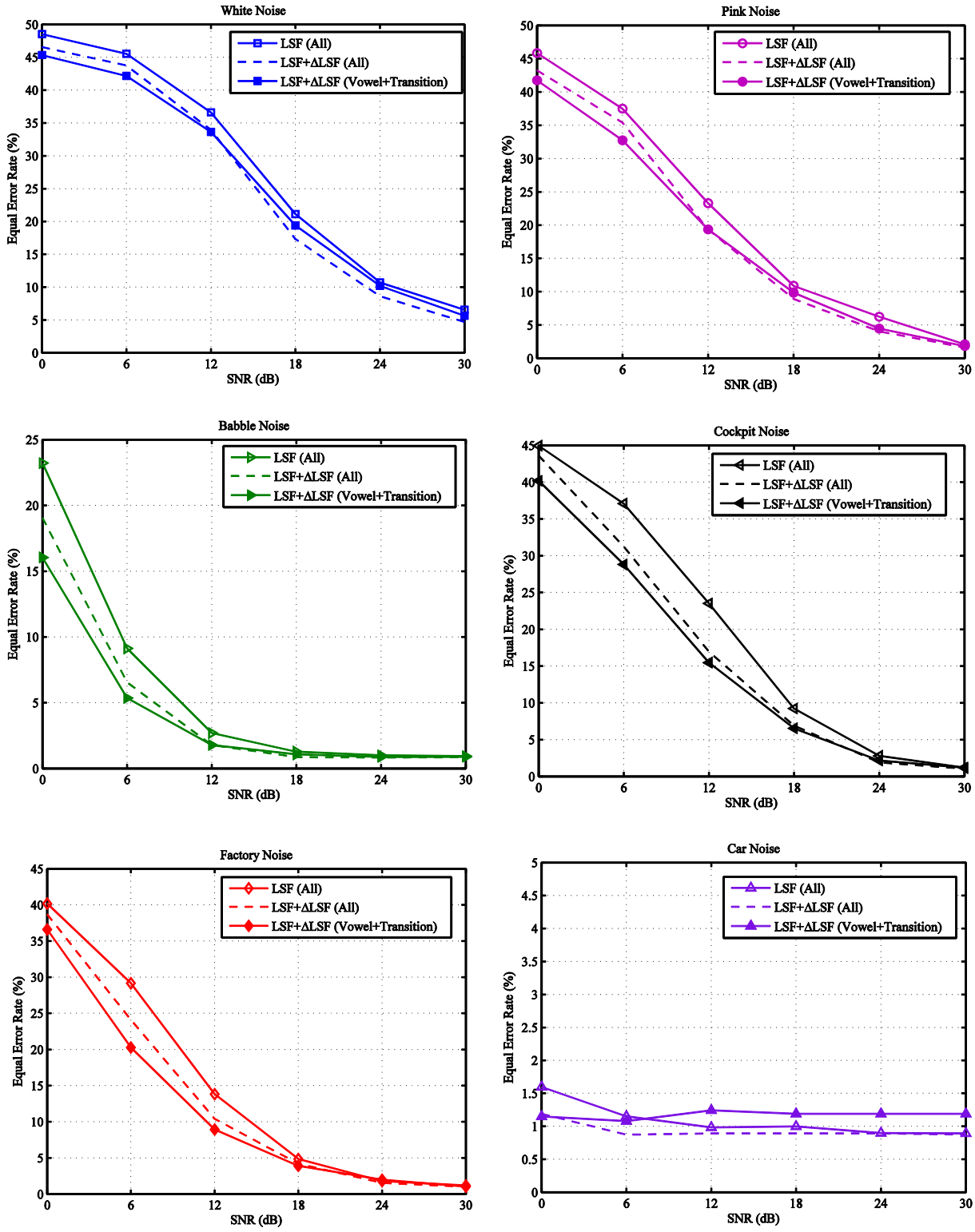

Fig. 7.8: Performance improvement over the baseline system by using score-level fusion and scoring exclusively on vowel and transition frames.

7.5 CONCLUSIONS

In this chapter, we tested the performance of our baseline LSF based SV system in clean as well as noisy conditions. By employing a score-level fusion of LSF and delta-LSF based classifiers, we were able to improve the noise-robustness of the SV system. In addition, we investigated whether utilizing only vowel and transition frames during verification improves the performance of the score-level fusion based SV system. The results are summarized in Table 7.8 and Fig. 7.8. In clean and high SNR scenarios, the best performance is obtained using the score-level fusion based system and scoring on all speech frames. However, scoring selectively on vowel and transition frames improves the SV performance in low SNR conditions. Thus, we propose a speaker verification system which fuses information from static LSF and dynamic ΔLSF based classifiers at the score-level, and utilizes a combination of vowel and transition zones of speech for scoring during the verification phase.

In the presence of white noise, the proposed system begins to provide the best performance from an SNR of 12 dB and below. At 12 dB SNR, the EER of the baseline system is 36.61%, whereas that of the proposed system is 33.63%. This corresponds to a relative EER decrease of 8.14%. Similarly, relative EER reductions of 6.54% and 7.44% are observed with respect to the baseline at SNRs of 0 dB and 6 dB respectively. Under pink noise, a relative EER decrease of 12.69% and 8.92% was observed at 6 dB and 0 dB SNR respectively. Similarly, in babble noise, a relative decrease of 30.89% and 41.29% was observed over the baseline EER at SNRs of 0 dB and 6 dB respectively.

Under cockpit noise, the proposed SV system begins to outperform the baseline from 18 dB SNR, at which there is a relative reduction of 8.52% in the EER. At 12 dB, 6 dB and 0 dB SNR, a relative EER decrease of 20.79%, 22.37% and 10.59% respectively is obtained, compared to the baseline. Similarly, in the presence of factory noise at 18 dB and 12 dB SNR, there is a relative decrease of 19.17% and 35.33% respectively in the EER. Relative EER reductions of 30.43% and 8.88% are observed at SNRs of 6 dB and 12 dB respectively.

For the SI system in Chapter 6, scoring only on vowel and transition frames improves SI performance even in high SNR conditions. Also, in the SI system, we were able to achieve more than 100% relative performance improvements over the baseline in some cases. However, in our SV system, the highest relative EER decrease is 41.29%.

These results suggest that frame-level selection is much more crucial in a speaker identification system than in a speaker verification system. Intuitively, this makes sense because speaker identification is a maximum likelihood problem, i.e. an N-class decision task, where N is the number of speakers. Thus, one corrupted frame score could affect the identification result significantly. However, speaker verification is a 2-class decision task and thus a few unreliable frames might not affect the verification performance significantly.

8 CONCLUSIONS AND FUTURE WORK

In this work, an automatic, text-independent speaker identification (SI) system and a speaker verification (SV) system were developed using Line Spectral Frequency (LSF) features. All experiments were conducted using the TIMIT speech corpus and the SPIB noise database. A simple Gaussian Mixture Model (GMM) framework was used for speaker enrollment in the SI system. The GMM was trained using k-means clustering and the Expectation-Maximization (EM) Algorithm. On the other hand, in the SV system, an adapted Gaussian Mixture Model framework was used for speaker enrollment. A Universal Background Model (UBM) was trained using a binary splitting procedure and the EM algorithm. The parameters of the UBM were used to create an adapted GMM for each enrolled speaker.

Under clean conditions, the baseline SI system had an identification accuracy of 99.7%, and the baseline SV system had an EER of 0.89%. The performance of the SI and SV systems were evaluated under white, pink, babble, cockpit, factory and car noise, at SNR levels ranging from 0-30 dB. A score-level fusion based technique was used to combine complementary information from static LSF and delta-LSF based classifiers. Score-level fusion was shown to improve the robustness of both the SI as well as the SV system under noisy conditions.

We also investigated the relative importance, or speaker-discriminative power of different speech zones under clean as well as noisy conditions. In particular, our hypothesis that rapidly varying transitions into and out of vowels are more speaker-discriminative than steady-vowel regions was tested. This analysis was performed during the identification/verification phase. In each test segment, the Vowel Onset Point (VOP) and Vowel End Point (VEP) locations were determined. Based on these locations, each test frame was classified as a vowel, transition or non-vowel frame. The discriminative power of each category was determined by evaluating the speaker recognition performance exclusively using frames of that category. Our results suggest that while transition regions are the most speaker-discriminative under high SNR conditions, high-energy vowel regions are most speaker-discriminative under low SNR conditions.

It was found that under relatively high SNR conditions, selectively utilizing frames of a particular category for speaker identification or verification did not result in any performance improvement. However, under noisy conditions, the performances of the proposed score-level fusion based SI and SV systems can be improved substantially by scoring exclusively on a combination of transition and vowel frames during the identification/verification phase.

During the course of this research, we did not investigate the effect of training speaker models using speaker-discriminative zones of speech due to lack of sufficient training data. The TIMIT database has only around 30 seconds of speech material per speaker. Thus, selecting only certain speech zones could impact the performance negatively, since the GMMs would not have sufficient training data.

Using a larger speaker database, we could explore the possibility of training speaker models using only transition regions of speech. It would be interesting to see whether this leads to improvements in performance.

Also, the algorithm used to localize transition zones of speech could be improved further, since the current algorithm detects only 80% of the vowel onset and end points. In addition, in the current setup, a fixed window around each VOP and VEP is considered as the transition region. A more noise-robust algorithm could be devised to accurately identify the transition regions. In addition, we could explore techniques to enhance the information present in transition zones under noisy conditions.

BIBLIOGRAPHY

[1] T. D. Ganchev, "Speaker recognition," University of Patras, 2005.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, pp. 12-40, 2010.

[3] *Tool Module: The Human Vocal Apparatus*. Available: http://thebrain.mcgill.ca/flash/capsules/outil_bleu21.html

[4] J. P. Campbell Jr, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437-1462, 1997.

[5] *How The Voice Works*. Available: http://www.entnet.org/content/how-voice-works

[6] *Vocal Tract Resonance*. Available: http://hyperphysics.phy-astr.gsu.edu/hbase/music/vocres.html

[7] D. A. Reynolds, "Automatic speaker recognition using Gaussian mixture speaker models," in *The Lincoln Laboratory Journal*, 1995.

[8] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225-254, 2000.

[9] *Speaker Recognition*. Available: http://en.wikipedia.org/wiki/Speaker_recognition

[10] N. Singh, R. Khan, and R. Shree, "Applications of Speaker Recognition," *Procedia Engineering*, vol. 38, pp. 3122-3126, 2012.

[11] T. Kinnunen, "Spectral features for automatic text-independent speaker recognition," *Licentiatesthesis, Department of computer science, University of Joensuu*, 2003.

[12] D. A. Reynolds, "Large population speaker identification using clean and telephone speech," *Signal Processing Letters, IEEE*, vol. 2, pp. 46-48, 1995.

[13] I. Pollack, J. M. Pickett, and W. H. Sumby, "On the Identification of Speakers by Voice," *The Journal of the Acoustical Society of America*, vol. 26, pp. 403-406, 1954.

[14] J. Shearme and J. Holmes, "An experiment concerning the recognition of voices," *Language and Speech*, vol. 2, pp. 123-131, 1959.

[15] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," *The Journal of the Acoustical Society of America*, vol. 35, pp. 354-358, 1963.

[16] S. Furui, "50 years of progress in speech and speaker recognition," *SPECOM 2005, Patras*, pp. 1-9, 2005.

[17] Q. Jin, J. Navratil, D. A. Reynolds, J. P. Campbell, W. D. Andrews, and J. S. Abramson, "Combining cross-stream and time dimensions in phonetic speaker recognition," in

Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, 2003, pp. IV-800-3 vol. 4.

[18] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 28, pp. 357-366, 1980.

[19] Q. Jin, "Overview of front-end features for robust speaker recognition," 2011.

[20] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task."

[21] K. S. Rao and S. Sarkar, *Robust Speaker Recognition in Noisy Environments*: Springer, 2014.

[22] W. Longbiao, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker recognition by combining MFCC and phase information in noisy conditions," *IEICE transactions on information and systems,* vol. 93, pp. 2397-2406, 2010.

[23] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining MFCC and phase information in noisy environments," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4502-4505.

[24] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America,* vol. 55, pp. 1304-1312, 1974.

[25] S. Ganapathy, J. Pelecanos, and M. K. Omar, "Feature normalization for speaker verification in room reverberation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 4836-4839.

[26] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally Weighted Linear Prediction Features for Tackling Additive Noise in Speaker Verification," *Signal Processing Letters, IEEE,* vol. 17, pp. 599-602, 2010.

[27] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America,* vol. 57, pp. S35-S35, 1975.

[28] K. Paliwal, "On the use of line spectral frequency parameters for speech recognition," *Digital signal processing,* vol. 2, pp. 80-87, 1992.

[29] C.-S. Liu, W.-J. Wang, M.-T. Lin, and H.-C. Wang, "Study of line spectrum pair frequencies for speaker recognition," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, 1990, pp. 277-280.

[30] I. V. McLoughlin, "Line spectral pairs," *Signal processing,* vol. 88, pp. 448-467, 2008.

[31] B. J. Lee, S. Kim, and H.-G. Kang, "Speaker recognition based on transformed line spectral frequencies," in *Intelligent Signal Processing and Communication Systems, 2004. ISPACS 2004. Proceedings of 2004 International Symposium on*, 2004, pp. 177-180.

[32] M. Sahidullah and G. Saha, "On the use of perceptual Line Spectral Pairs Frequencies for speaker identification," in *Communications (NCC), 2010 National Conference on*, 2010, pp. 1-5.

[33] H. Cordeiro and C. M. Ribeiro, "Speaker Characterization with MLSFs," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, 2006, pp. 1-4.

[34] F. Soong and A. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, 1986, pp. 877-880.

[35] J. S. Mason and X. Zhang, "Velocity and acceleration features in speaker recognition," in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, 1991, pp. 3673-3676 vol.5.

[36] K. Kumar, C. Kim, and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 4784-4787.

[37] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem, "Use of line spectral frequencies for emotion recognition from speech," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 3708-3711.

[38] T. Kinnunen and P. Alku, "On separating glottal source and vocal tract information in telephony speaker verification," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4545-4548.

[39] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, pp. 569-586, 1999.

[40] A. P. Lobo, "Glottal flow derivative modeling with the wavelet smoothed excitation," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, 2001, pp. 861-864 vol.2.

[41] B. LaRoy Berg and A. A. Beex, "Investigating speaker features from very short speech records," in *Circuits and Systems, 1999. ISCAS '99. Proceedings of the 1999 IEEE International Symposium on*, 1999, pp. 102-105 vol.3.

[42] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *the Journal of the Acoustical Society of America*, vol. 125, pp. 3289-3305, 2009.

[43] S. R. Mahadeva Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, pp. 1243-1261, 10// 2006.

[44] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *Signal Processing Letters, IEEE*, vol. 13, pp. 52-55, 2006.

[45] J. Wang and M. T. Johnson, "Vocal source features for bilingual speaker identification," in *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on*, 2013, pp. 170-173.

[46] W. M. Campbell, J. P. Campbell, T. P. Gleason, D. A. Reynolds, and W. Shen, "Speaker verification using support vector machines and high-level features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 2085-2094, 2007.

[47] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, pp. 1800-1803.

[48] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, *et al.*, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2003, pp. IV-784-7 vol. 4.

[49] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2003, pp. IV-788-91 vol. 4.

[50] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, pp. 782-796, 10// 2008.

[51] E. S. A. Stolcke, "The case for automatic Higher-Level features in forensic speaker recognition," 2008.

[52] D. Klusácek, J. Navratil, D. A. Reynolds, and J. P. Campbell, "Conditional pronunciation modeling in speaker detection," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2003, pp. IV-804-7 vol. 4.

[53] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[54] G. Tur, E. Shriberg, A. Stolcke, and S. Kajarekar, "Duration and Pronunciation Conditioned Lexical Modeling for Speaker Verification," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[55] F. K. Soong, A. E. Rosenberg, B.-H. Juang, and L. R. Rabiner, "Report: A vector quantization approach to speaker recognition," *AT&T technical journal*, vol. 66, pp. 14-26, 1987.

[56] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech communication*, vol. 17, pp. 91-108, 1995.

[57] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic Speaker Recognition with Support Vector Machines," in *Advances in Neural Information Processing Systems*, 2004, pp. 1377-1384.

[58] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, pp. 308-311, 2006.

[59] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, pp. 194-205, 1994.

[60] B. Xiang and T. Berger, "Efficient text-independent speaker verification with structural Gaussian mixture models and neural network," *Speech and Audio Processing, IEEE Transactions on,* vol. 11, pp. 447-456, 2003.

[61] A. K. Vuppala and K. S. Rao, "Speaker identification under background noise using features extracted from steady vowel regions," *International Journal of Adaptive Control and Signal Processing*, vol. 27, pp. 781-792, 2013.

[62] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[63] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 1711-1723, 2007.

[64] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and systems Magazine, IEEE*, vol. 11, pp. 23-61, 2011.

[65] L. Besacier and J. F. Bonastre, "Frame pruning for speaker recognition," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, 1998, pp. 765-768 vol.2.

[66] C.-S. Jung, M. YoungKim, and H.-G. Kang, "Selecting feature frames for automatic speaker recognition using mutual information," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, pp. 1332-1340, 2010.

[67] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech communication*, vol. 52, pp. 41-60, 2010.

[68] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, pp. 271-287, 2004.

[69] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 600-613, 2011.

[70] Y. Ma and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, pp. 1-18, 2013.

[71] J. Eatock and J. S. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, 1994, pp. I/133-I/136 vol. 1.

[72] E. G. Hansen, R. E. Slyh, and T. R. Anderson, "Speaker recognition using phoneme-specific GMMs," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[73] S. S. Kajarekar and H. Hermansky, "Speaker verification based on broad phonetic categories," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[74] N. Fatima and T. F. Zheng, "Syllable category based short utterance speaker recognition," in *Audio, Language and Image Processing (ICALIP), 2012 International Conference on*, 2012, pp. 436-441.

[75] S. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 2552-2565, 2011.

[76] G. Pradhan and S. R. M. Prasanna, "Speaker Verification by Vowel and Nonvowel Like Segmentation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 854-867, 2013.

[77] N. Fatima, S. Aftab, R. Sultan, S. A. H. Shah, B. M. Hashmi, A. Majid, *et al.*, "Speaker recognition using lower formants," in *Multitopic Conference, 2004. Proceedings of INMIC 2004. 8th International*, 2004, pp. 125-130.

[78] Available: http://clas.mq.edu.au/speech/acoustics/speech_spectra/fft_lpc_settings.html

[79] J. Louradour, K. Daoudi, R. André-Obrecht, and P. Sabatier, "Discriminative power of transient frames in speaker recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005, pp. 613-616.

[80] J. Louradour, R. André-Obrecht, and K. Daoudi, "Segmentation and relevance measure for speaker verification," in *INTERSPEECH*, 2004.

[81] C.-S. Jung, K. J. Han, H. Seo, S. S. Narayanan, and H.-G. Kang, "A variable frame length and rate algorithm based on the spectral Kurtosis measure for speaker verification," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[82] S. Prasanna, B. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, pp. 556-565, 2009.

[83] S. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation information," in *INTERSPEECH*, 2005, pp. 1133-1136.

[84] B. D. Sarma, S. S. Prajwal, and S. Mahadeva Prasanna, "Improved Vowel Onset and offset points detection using bessel features," in *Signal Processing and Communications (SPCOM), 2014 International Conference on*, 2014, pp. 1-6.

[85] A. Nayeemulla Khan and B. Yegnanarayana, "Vowel onset point based variable frame rate analysis for speech recognition," in *Intelligent Sensing and Information Processing, 2005. Proceedings of 2005 International Conference on*, 2005, pp. 392-394.

[86] J. D. Markel and A. H. J. Gray, *Linear Prediction of Speech*: Springer Berlin Heidelberg, 2013.

[87] *Voice Acoustics*. Available: http://newt.phys.unsw.edu.au/jw/voice.html

[88] K. M. M. Prabhu, *Window Functions and Their Applications in Signal Processing*: CRC Press, 2013.

[89] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7229-7233.

[90] *Linear Prediction of Speech*. Available: http://research.cs.tamu.edu/prism/lectures/sp/l7.pdf

[91] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561-580, 1975.

[92] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, pp. 1419-1426, 1986.

[93] F. Zheng, Z. Song, L. Li, W. Yu, F. Zheng, and W. Wu, "The distance measure for line spectrum pairs applied to speech recognition."

[94] *TIMIT Database*. Available: https://catalog.ldc.upenn.edu/LDC93S1

[95] *SPIB Noise Dataset*. Available: http://spib.linse.ufsc.br/noise.html

[96] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*, ed: Springer, 2009, pp. 659-663.

[97] *MSR Identity Toolkit*. Available: http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-11/IdentityToolbox/

[98] 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.

[99] M. Faundez-Zanuy and E. Monte-Moreno, "State-of-the-art in speaker recognition," *Aerospace and Electronic Systems Magazine, IEEE*, vol. 20, pp. 7-12, 2005.

[100] B. Yegnanarayana and S. V. Gangashetty, "Epoch-based analysis of speech signals," *Sadhana*, vol. 36, pp. 651-697, 2011.

[101] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, pp. 1602-1613, 2008.