Characterization of the voice source by the DCT for speaker information

A Thesis

Submitted For the Degree of

Master of Science (Engineering)

in the Faculty of Engineering

by

Abhiram B

(Under the guidance of Prof. A. G. Ramakrishnan)

Department of Electrical Engineering

Indian Institute of Science

BANGALORE – 560 012

MAY 2014

TO

My parents

for

Everything they have given me

Acknowledgements

I consider myself blessed to have worked in the serene atmosphere of IISc, amidst great scientific minds of our country. Over the past two and a half years, I have learnt a lot of things here, which have had a profound impact on my thinking. I take this opportunity to express my thanks to those who were responsible for it, and to all those who have helped me in my endeavour.

First, I wholeheartedly thank Prof. A. G. Ramakrishnan for all the encouragement, ideas and help he gave me. He listened patiently to anything I told him and was always free for a discussion. Working with him has given me a certain discipline in the way of doing things, which I am sure will stay with me forever.

If you ask me to name a person from whom I have learnt the most about speech, it is Dr. T. V. Ananthapadmanabha. His insights into speech science are amazing, and he has shown me how to be a successful researcher and entrepreneur at the same time. I feel proud to have gained a thing or two from his vast knowledge and experience.

I am fortunate to have worked with Dr. S. R. Mahadeva Prasanna of IITG, who inspired me from his work ethics, and made my stay at IITG enjoyable and rewarding. I thank him for his suggestions, which have improved the quality of this work. I also thank Prof. Sastri, Prof. K. R. Ramakrishnan, Prof. T. V. Sreenivas, Dr. Chandra Sekhar Seelamantula and Dr. Chandra R. Murthy, who made coursework a pleasure.

I am blessed to be the disciple of Prof. Dr. C. A. Shreedhar, who amazes me with his passion for music, and with his beautiful ragaalaapanas and swaras. I am sure that the positive vibrations I gain by learning music from him will have a great impact on my life.

"God could not be everywhere so he made Mothers" - goes a popular saying. I cannot imagine my life without Amma, who has dedicated her life to my family, and Appa, who has

given me everything I have asked him.

A world without friends is no world at all. My heartfelt thanks to all my IISc buddies, Prathosh, Pramod, Chenna, Suraj, Vishwas, Pai, Rama,...the list goes on. The late night talks, discussions on music, sports, philosophy, literature, movies... all have collectively given a new dimension to my way of living.

I have gained a lot from the company of my friends outside IISc. I especially thank Praveen, Shravan and Kshiteesh, who are my friends for life. A special thanks with love goes to my fiance Arpita, who has made life more beautiful since I met her.

I thank my Doddamma and Doddappa for treating me like their own son, and helping me with all the things during my stay here. I thank Sanju for his humour and all the discussions on almost everything in the world, and Shashu for giving me great company at home.

I express my thanks to Vijay, Nazreen, Rajaram, Harshitha and all the members of MILE Lab for their well-wishes, and to Suresh, Haris, Nagaraj and Ramesh, who made my time in IITG memorable.

My sincere thanks goes to Mr. Channegowda, Mr. Purushottam and all the department office members for their help in official matters. Finally, I thank everyone else whose names I have by mistake missed to mention here.

Abstract

Extracting speaker-specific information from speech is of great interest to both researchers and developers alike, since speaker recognition technology finds application in a wide range of areas, primary among them being forensics and biometric security systems.

Several models and techniques have been employed to extract speaker information from the speech signal. Speech production is generally modeled as an excitation source followed by a filter. Physiologically, the source corresponds to the vocal fold vibrations and the filter corresponds to the spectrum-shaping vocal tract. Vocal tract-based features like the mel-frequency cepstral coefficients (MFCCs) and linear prediction cepstral coefficients have been shown to contain speaker information. However, high speed videos of the larynx show that the vocal folds of different individuals vibrate differently. Voice source (VS)-based features have also been shown to perform well in speaker recognition tasks, thereby revealing that the VS does contain speaker information. Moreover, a combination of the vocal tract and VS-based features has been shown to give an improved performance, showing that the latter contains supplementary speaker information.

In this study, the focus is on extracting speaker information from the VS. The existing techniques for the same are reviewed, and it is observed that the features which are obtained by fitting a time-domain model on the VS perform poorly than those obtained by simple transformations of the VS. Here, an attempt is made to propose an alternate way of characterizing the VS to extract speaker information, and to study the merits and shortcomings of the proposed speaker-specific features.

The VS cannot be measured directly. Thus, to characterize the VS, we first need an estimate of the VS, and the integrated linear prediction residual (ILPR) extracted from the speech signal is used as the VS estimate in this study. The voice source linear prediction

model, which was proposed in an earlier study to obtain the ILPR, is used in this work.

It is hypothesized here that a speaker's voice may be characterized by the relative proportions of the harmonics present in the VS. The pitch synchronous discrete cosine transform (DCT) is shown to capture these, and the gross shape of the ILPR in a few coefficients. The ILPR and hence its DCT coefficients are visually observed to distinguish between speakers. However, it is also observed that they do have intra-speaker variability, and thus it is hypothesized that the distribution of the DCT coefficients may capture speaker information, and this distribution is modeled by a Gaussian mixture model (GMM).

The DCT coefficients of the ILPR (termed the DCTILPR) are directly used as a feature vector in speaker identification (SID) tasks. Issues related to the GMM, like the type of covariance matrix, are studied, and it is found that diagonal covariance matrices perform better than full covariance matrices. Thus, mixtures of Gaussians having diagonal covariances are used as speaker models, and by conducting SID experiments on three standard databases, it is found that the proposed DCTILPR features fare comparably with the existing VS-based features. It is also found that the gross shape of the VS contains most of the speaker information, and the very fine structure of the VS does not help in distinguishing speakers, and instead leads to more confusion between speakers. The major drawbacks of the DCTILPR are the session and handset variability, but they are also present in existing state-of-the-art speaker-specific VS-based features and the MFCCs, and hence seem to be common problems. There are techniques to compensate these variabilities, which need to be used when the systems using these features are deployed in an actual application.

The DCTILPR is found to improve the SID accuracy of a system trained with MFCC features by 12%, indicating that the DCTILPR features capture speaker information which is missed by the MFCCs. It is also found that a combination of MFCC and DCTILPR features on a speaker verification task gives significant performance improvement in the case of short test utterances. Thus, on the whole, this study proposes an alternate way of extracting speaker information from the VS, and adds to the evidence for speaker information present in the VS.

Contents

Acknowledgements **i**

Abstract **iii**

1 Introduction **1**
1.1 The source-filter model of speech production 1
1.1.1 Human speech production 1
1.1.2 The source-filter model 2
1.2 Speaker recognition 4
1.3 Motivation to use VS-based features for SID 6
1.4 Objectives of the thesis 7

2 Literature survey **8**
2.1 Glottal inverse filtering 8
2.1.1 Linear prediction and closed-phase analysis 10
2.2 Review of VS-based features for speaker recognition 12
2.2.1 LF model parameters 13
2.2.2 Voice source cepstral coefficients 16
2.2.3 Deterministic plus stochastic model (DSM) of the LP residual 17
2.2.4 Parametric vs non-parametric characterizations 19

3 Characterization of the voice source by the DCT **20**
3.1 Revisiting the source-filter model - the voice source linear prediction model . 20
3.1.1 The integrated linear prediction residual 21
3.2 The DCT of ILPR: A characterization of the VS 25
3.2.1 Motivation for using the DCT to characterize the VS 25
3.2.2 Pitch synchronous analysis 26
3.2.3 Obtaining the pitch synchronous DCT of ILPR 28
3.2.4 Pre-processing used to obtain the DCT of ILPR 29
3.3 The DCT of ILPR as a feature vector for SID 31
3.3.1 Variability of the ILPR 31
3.3.2 Number of DCT coefficients 35

4 Experiments and results **40**
4.1 Gaussian mixture models for SID 40
4.2 Issues with GMMs - The covariance matrix type and the number of Gaussians 41
4.2.1 Full covariance matrices 42
4.2.2 Diagonal covariance matrices 45
4.3 Description of SID experiments 47
4.3.1 Databases, training and test data 47
4.4 The number of DCT coefficients for a 'good' feature vector 48
4.5 Results and discussion 50
4.5.1 Results on TIMIT 50
4.5.2 Results on YOHO 50
4.5.3 Results on NIST 2003 52
4.5.4 Results with the DFT of ILPR 55
4.6 Speaker verification experiments on short test utterances 56
4.6.1 The baseline i-vector system 57
4.6.2 Performance of DCTILPR features 58
4.6.3 Combination of MFCC and DCTILPR features 58

5 Conclusion and future work **61**

A Interpretation of the DCT coefficients as harmonics **64**

B Resynthesis of the speech signal by varying the number of retained DCT coefficients of the ILPR **66**
B.1 Resynthesis of speech from ILPR by varying M 66

Bibliography **70**

Publications based on this thesis **77**

List of Tables

2.1 SID performance of LF model parameters on the TIMIT test set 15
2.2 SID performance of parametric and non-parametric features 19

4.1 SID performance on the TIMIT test set using GMMs with full covariance matrices, with the optimum G determined in different ranges using cross-validation. The range 5-10 gives the best result. 45
4.2 Comparison of SID performance of different VS-based features on the 168-speaker TIMIT test set. GMMs with 16 Gaussians having diagonal covariance matrices are employed as the speaker models. 50
4.3 Percentage of speakers classified in different positions with DCTILPR (with DSM) when test data is taken from different recording sessions in the 138-speaker YOHO database. GMMs with 16 Gaussians having diagonal covariance matrices are employed as the speaker models. 51
4.4 SID accuracies of the DCTILPR, DSM and VSCC features on the 138-speaker YOHO database, averaged across the 4 sessions. GMMs with 16 Gaussians having diagonal covariance matrices are employed as the speaker models. . . 52
4.5 Comparison of SID performance of DCTILPR with MFCC features and their combination on the 110-speaker subset of the NIST 2003 database under the same and different handset conditions. 53
4.6 Performance comparison of DFTILPR and DCTILPR 56
4.7 Results of the baseline i-vector system on the entire 356-speaker NIST 2003 dataset using MFCC features for limited duration test segments. 58
4.8 Results of the i-vector system on the entire 356-speaker NIST 2003 database using DCTILPR features for limited duration test segments. 59
4.9 Performance of the proposed i-vector system for short test segments on the 356-speaker NIST 2003 database fusing DCTILPR and MFCC-trained classifiers at the score level. Improvement of the performance metrics over the baseline system are also listed. 60

List of Figures

1.1 Human speech production apparatus (taken from [1]) 2
1.2 The source-filter model and two views of the voice source 3
1.3 A typical speaker recognition system. The speaker claim and the binary decision (accept/reject) are valid only in the case of speaker verification. 5

2.1 Glottal flow/volume velocity (top panel) and its derivative (bottom panel) (taken from [2]) 9
2.2 Block diagram of the feature extraction process developed by Plumpe et. al. (taken from [3]) 13
2.3 A single glottal cycle with the LF model parameters corresponding to different phases (taken from [3]) 14
2.4 Fine structure of the VS showing aspiration and ripple (taken from [3]) . . . 15
2.5 Block diagram of the method to extract VSCCs (taken from [4]) 16
2.6 Block diagram of the method to extract the deterministic and stochastic components of the LP residual (taken from [5]) 18

3.1 VSLP analysis model along with spectra at each stage - the ILPR is an estimate of the VS (taken from [6]) 22
3.2 Difference between usual LP analysis and VSLP analysis - The LPR and the ILPR 23
3.3 ILPR of synthetic speech. (a) A synthesized vowel segment. (b) The VS pulse train used for synthesizing the vowel (blue) and the estimated ILPR (red). . 24
3.4 (a) ILPR estimated from a 20 ms segment of a synthetic vowel and (b) its cyclically shifted version; (c) and (d): The first 20 DCT coefficients of the signals in (a) and (b), respectively. 27
3.5 Block diagram of the method to extract the pitch synchronous DCT of ILPR 28
3.6 A TIMIT speech segment with the ILPR and the estimated epoch locations; the epochs can be seen at the negative peaks of the ILPR 29
3.7 A segment from a TIMIT utterance and the corresponding MNCC plot overlaid on it; observe that MNCC>0.6 for voiced phones and <0.6 for unvoiced phones. 31
3.8 ILPR and its DCT from different phonetic contexts in voiced segments from three different utterances of a single TIMIT speaker. The shapes of the ILPRs are similar and so are the DCT coefficient vectors. 32

3.9 ILPR and its DCT from the vowel /a/ in the word 'wash' for three TIMIT speakers. The ILPR shape varies from speaker to speaker and the DCT coefficients vary with waveform shape. 33
3.10 2-D scatter plot of DCT coefficients 5 vs 10 of the ILPR cycles from a TIMIT speaker. Cluster boundaries are roughly indicated by ellipses. 34
3.11 3-D scatter plot of DCT coefficients 1 vs 5 vs 10 of the ILPR cycles from a TIMIT speaker. Cluster boundaries are roughly indicated by ellipses. 34
3.12 3-D scatter plot of DCT coefficients 1 vs 5 vs 10 of the ILPR cycles from another TIMIT speaker. There is only one prominent cluster and the data distribution is different from that of the previous speaker. 35
3.11 One period of ILPR and its reconstruction from its truncated DCT, varying the number of DCT coefficients retained. 38
3.12 Mean of the percentage of signal energy captured as a function of M, computed using data from 100 TIMIT speakers. 39

4.1 Test phase of the SID system using GMM speaker models 42
4.2 Cross-validation strategy to determine the optimal number of Gaussians for modeling each speaker. $\lambda_{i,j}$ represents the GMM with j Gaussians for speaker i. 43
4.3 Histograms of training samples from two TIMIT speakers assigned to various Gaussian components of the GMMs for different G. 44
4.4 SID accuracy on the 168-speaker TIMIT test set versus the number of Gaussians (chosen to be the same for all speakers) with diagonal covariance matrices. 46
4.5 SID accuracy versus the number of DCT coefficients (M) on the 168-speaker TIMIT test set and the 138-speaker YOHO database. GMMs with 16 Gaussians having diagonal covariance matrices are employed as the speaker models. 49
4.6 SID performance of the combination of DCTILPR and MFCC-trained classifiers as a function of α for the same handset condition on the 110-speaker subset of the NIST 2003 database. 54
4.7 SID performance of the combination of DCTILPR and MFCC-trained classifiers as a function of α for the different handset condition on the 110-speaker subset of the NIST 2003 database. 54
4.8 Block diagram of baseline i-vector system (taken from [7]) 58
4.9 EER and DCF versus α for the fusion of the DCTILPR and MFCC-trained classifiers on the 356-speaker NIST 2003 database. 60

A.1 Relation between DCT and DFT of the even symmetric extension 64
A.2 Relation between DFT and DTFS of the periodicized version 65

B.-1 Voiced speech signal and its reconstruction from the truncated DCT of the ILPR, for varying number of DCT coefficients retained. The resynthesis error energy (as a percentage of the original signal energy) is also given. 69

Chapter 1

Introduction

1.1 The source-filter model of speech production

Speech is the most widely used form of communication by humans. It is a very complex code which has many types of information, e.g., phonetic information, prosody and emotion, speaker information, language and meaning. The human speech perception mechanism can decode these information, and as engineers, we are interested in emulating it. However, to accomplish this task, all we have is a signal from a transducer. In order to analyze and make sense out of it, we need to model the speech production mechanism, and for this the source-filter model [8] is used.

1.1.1 Human speech production

The human speech apparatus is shown in Figure 1.1 along with an enlarged cross-sectional view of the vocal folds, which are muscular tissues in the larynx. The slit between the vocal folds is the glottis, and the vocal folds vibrate, causing changes in the glottal area. Depending on the production mode, speech can be classified into two categories:

1. Voiced speech – Here, the vocal folds vibrate quasi-periodically. The air from the lungs passes through the vibrating vocal folds and then passes through the vocal tract. Thus, voiced speech (e.g., vowels, semivowels and glides) has a quasi-periodic nature, and has harmonics associated with it.

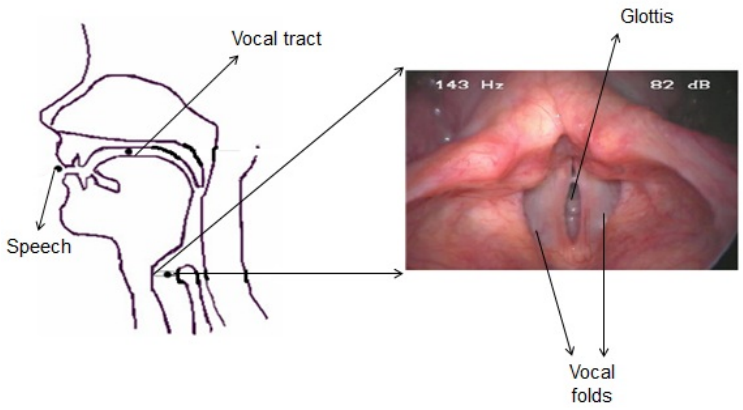


Figure 1.1: Human speech production apparatus (taken from [1])

2. Unvoiced speech – Here, the vocal folds do not vibrate, but are held close together, causing a turbulent flow of air. Unvoiced speech (e.g., fricatives like /s/ and /f/, stops like /k/ and /p/) is non-periodic and noise-like.

After passing through the glottis, the air passes through the vocal tract, and is radiated from the lips as speech. The vocal tract can be constricted at various points, leading to the production of different phones.

1.1.2 The source-filter model

Figure 1.2 illustrates the source-filter model. Speech is modeled as the output of a filter (representing the vocal tract) which is excited by a source (representing the airflow through the glottis). The source is modeled as a random signal generator for unvoiced speech. For voiced speech, the source is termed the voice source (VS). There are two views of the VS:

1. Impulse train – This is a simplified view of the VS. As shown in Figure 1.2 (red plot), since an impulse train has a flat spectrum, all the spectral variation in speech is attributed to the filter, which has several peaks in its magnitude response representing the resonances of the vocal tract.

2. Glottal pulses – This is a detailed view of the source. The glottal airflow changes during each cycle based on the movement of the vocal folds, and this variation is considered as the source, which excites a vocal tract filter having a magnitude response

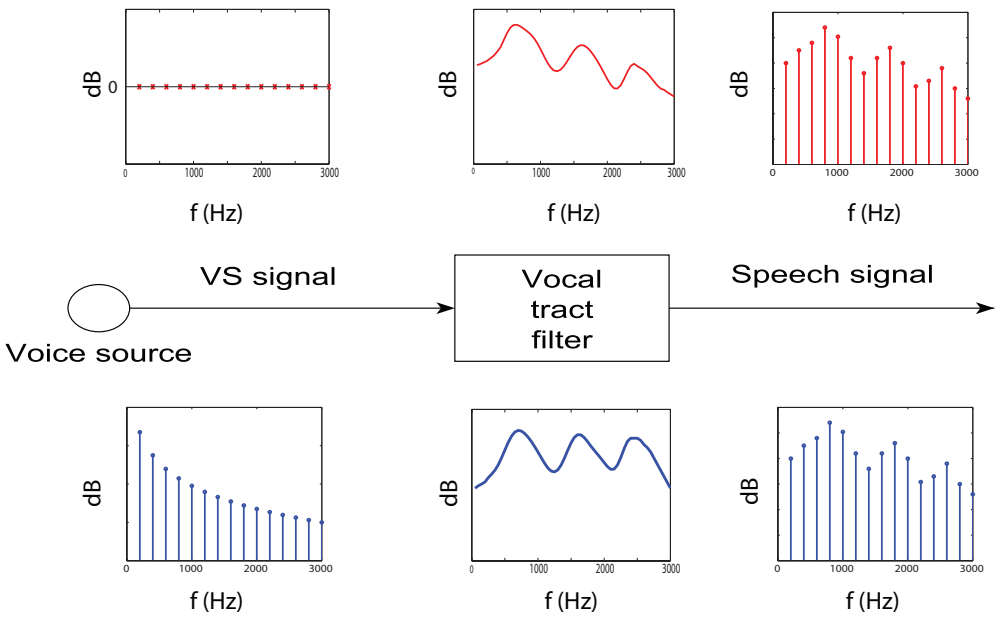


Figure 1.2: The source-filter model and two views of the voice source

as shown in Figure 1.2 (blue plot). We can see that the glottal pulse train typically has a harmonic spectrum with a spectral tilt, while the filter's magnitude response has peaks at resonances of the vocal tract, but does not have a significant spectral tilt. Thus, most of the spectral tilt in the spectrum of the speech signal can be attributed to the source.

Irrespective of how we view the VS, the VS signal is the input to the vocal tract filter with speech as the output. Thus, there is a convolutive relationship between the VS, the impulse response of the filter, and the speech signal. Specifically,

$$u_g(n) * h(n) = s(n) \tag{1.1}$$

where $u_g(n)$ is the VS, which is the glottal flow derivative (the reason for considering the derivative and not the flow itself as the VS will be explained later), $h(n)$ is the impulse response of the filter, and $s(n)$ is the speech signal. In the frequency domain, this translates to a multiplication of the respective Fourier transforms, i.e,

$$U_g(e^{j\omega}).H(e^{j\omega}) = S(e^{j\omega}) \tag{1.2}$$

Since the VS is quasi-periodic, it has a harmonic spectrum, and with the log-magnitude scale, this is added to the filter's frequency response to get the spectrum of the speech signal, as shown in Figure 1.2.

An implicit assumption in the source-filter model is that the source and the filter are independent of each other. In reality, there are interactions between the two, as shown by many studies [9, 10]. But, in this work, we assume that these source-filter interactions are negligible, since most of the previous studies have applied the independent source-filter model for many applications successfully [11, 12, 6].

Also, in this work, we are concerned only with the VS, and hence we ignore the unvoiced speech regions during processing. We shall revisit the source-filter model in greater detail in later sections.

1.2 Speaker recognition

Speaker recognition is the task of identifying the speaker (the one who spoke) from the spoken utterance. Humans do this remarkably well - we can recognize who is speaking without looking at the person even when he/she is speaking from a telephone, and also in the presence of noise and multiple other speakers. The goal of automatic speaker recognition[1] is to emulate this with a machine.

A block diagram of a typical speaker recognition system is given in Figure 1.3 [13]. Such a system can operate in two modes:

1. Speaker identification (SID)– In this mode, the system is presented with a test utterance and it has to identify it as belonging to one among several speakers.

2. Speaker verification – In this mode, the system is presented with a test utterance and a claim that it belongs to a particular speaker, and it has to verify whether the claim is right or wrong.

[1]Henceforth, in this thesis, speaker recognition refers to automatic speaker recognition, unless specified otherwise.

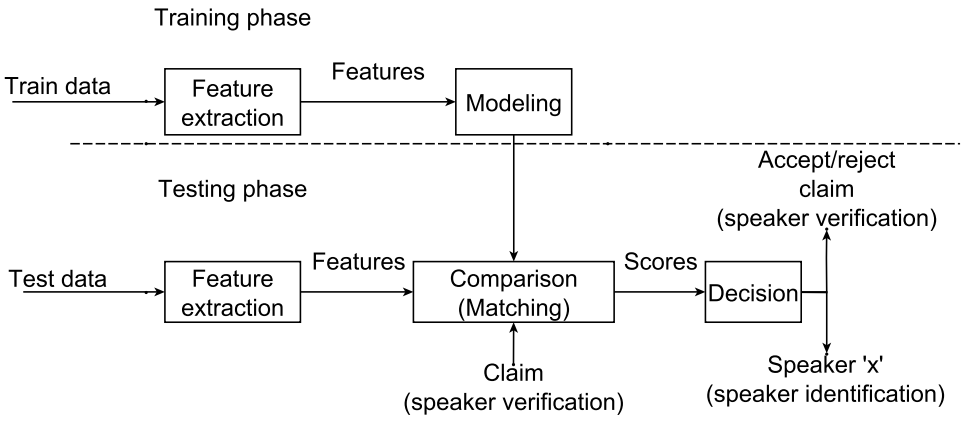


Figure 1.3: A typical speaker recognition system. The speaker claim and the binary decision (accept/reject) are valid only in the case of speaker verification.

One can see from Figure 1.3 that, in a speaker recognition task, the first phase is the training phase, during which models are learnt for each speaker using features extracted from training data (usually few minutes of read speech or extracts of one side of a conversation). The model is usually a probabilistic model like the Gaussian mixture model (GMM) or a discrete codebook-based vector quantization (VQ) model. The second phase is the testing phase, during which the features extracted from the test data (a few seconds of speech) are matched (compared) with the learnt models to obtain the scores and finally, a rule-based decision is made.

Note that the feature extraction block is the same for both the training and the test phases, and is the very first step in any speaker recognition system. As the name says, this block (hopefully) extracts speaker-specific features from the speech signal. This step is crucial in speaker recognition, since the modelling and comparisons depend upon what features we extract from the speech signal. In general, a good feature for speaker recognition is one which has small intra-speaker variability and large inter-speaker variability.

Speaker recognition systems can be classified into two types:

1. Text-dependent – During the training phase, each speaker should utter a phrase to enroll himself. During the testing phase, the user has to speak the same phrase as was spoken during the training, and the system matches the train and test patterns and gives the result.

2. Text-independent – There is no restriction on training and testing speech. The user can speak any phrase and the system tries to recognize the speaker by extracting features which change little/do not change across different phones.

Text-dependent speaker recognition can be applied to verify the account holder in a bank ATM, to check a person's identity in a domestic electronic security system, etc. The major application of text-independent speaker recognition is in forensics, where we would have recordings from a crime scene and several suspects, out of whom we have to identify who committed the crime. In this work, we focus on text-independent SID[2].

1.3 Motivation to use VS-based features for SID

The VS signal contains information which can be used in many applications. For speech synthesis, the VS pulses synthesized by time-domain models have been used [14, 15]. Also, the VS estimated from the speech signal has been used as the source signal to produce natural-sounding speech [16, 17]. Another application where the VS is useful is in voice pathology [18, 19], where it helps us to study dysfluencies and other vocal disorders. Further, studies show that the shape of the VS pulse influences perceived voice quality [6, 20].

High speed videos of the vocal folds indicate that their vibrations are different for different speakers [21]. This may be due to factors like vocal fold shape, size and position, which may change from person to person since each of us have our own unique anatomy. Thus, if we view the VS as a glottal pulse train, we may be able to distinguish speakers based on the differences in their glottal pulses. Accordingly, the glottal pulse view, which is closer to the actual physiology, is assumed in this work.

There are previous studies which show that VS-based features can be used successfully in SID tasks [22, 3, 4, 5]. This is another motivation to mine the VS for speaker information.

Since we assume negligible source-filter interactions, the VS-based features are expected not to change much with the vocal tract configuration, and hence the VS-based features of a particular speaker may remain fairly constant irrespective of the phonetic composition of

[2]Henceforth, in this thesis, SID refers to text-independent SID, unless specified otherwise.

the utterance spoken. Such features are particularly attractive for text-independent SID, and this is another motivation for us to explore VS-based features.

1.4 Objectives of the thesis

The objectives of this thesis are:

1. To explore the existing VS-based speaker-specific features and to propose a new characterization of the VS leading to a new VS-based speaker-specific feature.

2. To evaluate the SID performance of the proposed feature on different databases and infer its merits and demerits.

Here, it must be said that there are vocal-tract or *system*-based features which capture speaker information. Probably the most successful and widely used among such features are the mel-frequency cepstral coefficients (MFCCs). These capture the energies of the mel filterbank outputs of the magnitude spectrum of the speech signal via its cepstral coefficients, which in essence captures the speech spectral envelope, and hence the MFCCs are predominantly vocal tract or system-based features. However, in this work, we focus on VS-based features for the motivational reasons mentioned in the previous section. We would like to explicitly state that the focus of this work is neither speaker recognition per se nor speaker modeling techniques. So, in the next chapter, we only review the VS-based features in the literature.

Chapter 2

Literature survey

2.1 Glottal inverse filtering

The vocal fold vibrations are difficult to measure directly. It requires invasive techniques like high-speed stroboscopy (in which a strobe with a camera and a light source is inserted into the pharynx and high-speed videos, with frame rates of the order of 1000 frames/second, are obtained) to capture these vibrations, but even such techniques capture only the change in glottal area as a function of time and not the actual glottal flow (or volume velocity) [2]. There are attempts to relate the glottal flow with the area function [23], but they still seem to be inconclusive. Moreover, in a number of applications where the VS is used, we cannot measure these vibrations directly. For example, in a forensic speaker recognition task, we cannot have stroboscope data from the crime scene. Even where they are possible, these invasive techniques are time consuming and are uncomfortable for the user.

For the reasons mentioned above, the vocal fold vibrations are not measured directly, but are estimated from the speech signal itself. Obviously, we need a model to estimate them, and the source-filter model explained in section 1.1.2 is used to this end. The generic technique used to estimate the VS from the speech signal is called glottal inverse filtering (GIF). This is because these methods involve 'canceling' the effect of the vocal tract by passing the speech signal through a filter having a transfer function which is the inverse of the vocal tract filter's transfer function.

To understand how various GIF methods work, we need to take a look at the time-domain

glottal flow and derivative waveforms. Figure 2.1 shows a typical glottal flow waveform (top panel) along with its derivative (bottom panel). We can see that there are different phases in a glottal cycle:

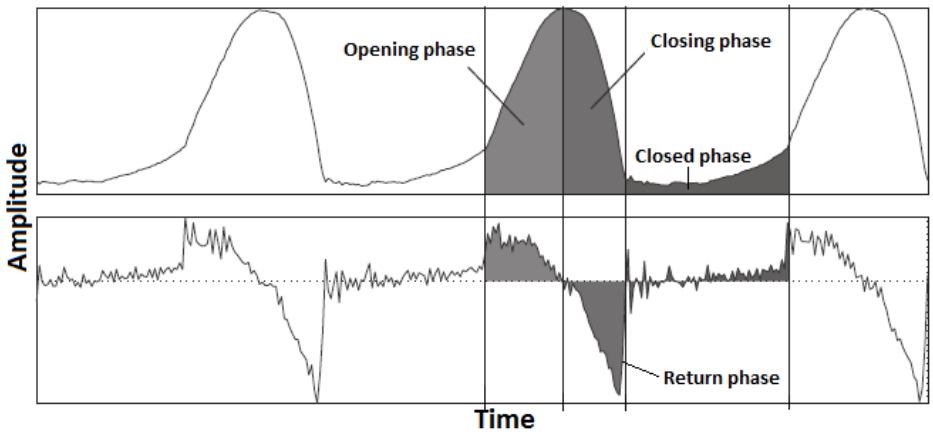


Figure 2.1: Glottal flow/volume velocity (top panel) and its derivative (bottom panel) (taken from [2])

1. Open phase – this is the phase when the vocal folds are open. It can be subdivided into two phases:

   - Opening phase – when the vocal folds begin to open, the glottal flow increases steadily and at the instant the vocal fold opening is maximum, the flow reaches a maximum. The flow derivative is positive throughout this phase; it increases, reaches a maximum and comes back to zero at the instant of maximum flow.
   - Closing phase – the vocal folds begin to close, and the flow tends to zero. The flow derivative is negative throughout this phase, and, at an instant when the vocal folds are almost closed, it reaches a negative maximum, and this instant is called the glottal closure instant (GCI).

2. Return phase – after the closing phase, the vocal folds are not yet completely closed, and they do not close suddenly. After the instant when the flow derivative is maximum, the vocal folds take a short duration of time to close fully, and this is reflected in the flow derivative signal as a sharp transition from the negative peak to zero.

3. Closed phase – this is the phase when the vocal folds are closed. The flow and its derivative are zero throughout this phase. However, as shown in Figure 2.1, sometimes and for some speakers, the vocal folds do not close fully and there will be some residual flow.

GIF has a long history dating back to the 1950's [2]. Miller's pioneering work [24] used lumped elements in an electrical network to cancel out the effect of the vocal tract. Later studies were conducted by Gunnar Fant, Gauffin, Rothenberg and others, and Rothenberg's study [25] is especially notable since it involved a mask to record the volume velocity at the mouth and inverse filter this signal to get the glottal volume velocity. In all methods till this time, the implementation was analog. In the 1970's, a method called the closed-phase covariance method [11] was introduced, which was the first digital GIF implementation. In this method, the vocal tract is modeled as an all-pole digital filter, and covariance method of linear prediction (LP) analysis is applied in the closed-phase of the glottal cycle to estimate the filter coefficients. General LP analysis and closed-phase LP analysis are described below.

2.1.1 Linear prediction and closed-phase analysis

To put it simply, LP analysis tries to estimate a speech signal sample as a linear combination of a finite number of past samples. If $s(n)$ is the speech signal and p past samples are used for estimation (called the prediction order), then

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + e(n) \tag{2.1}$$

where $e(n)$ is the estimation error, also called the LP residual. In the z-domain, equation 2.1 can be written as

$$S(z) = S(z)\sum_{k=1}^{p} a_k z^{-k} + E(z) \Longrightarrow S(z) = \frac{E(z)}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.2}$$

Note that, since the vocal tract configuration changes continuously with time, the vocal tract filter is time-varying. But, for analysis purposes, we consider the vocal tract filter to

slowly vary with time. In other words, we consider the speech signal to be quasi stationary, i.e, stationary for a short interval of time, of the order of a few milliseconds. Thus, LP analysis assumes quasi stationarity. Now, from our source-filter model equation (equation 1.1), we have

$$S(z) = U_g(z).H(z) \tag{2.3}$$

Comparing equations 2.2 and 2.3, if the vocal tract filter is modeled as an all-pole filter, i.e, if $H(z) = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}}$, then $U_g(z) = E(z) \Longrightarrow u_g(n) = e(n)$. Now, we know that, in the closed-phase of the glottal cycle, assuming complete closure of the vocal folds, $u_g(n) = 0$. Thus, the coefficients a_k (called the LP coefficients) can be estimated in the closed-phase by minimizing $\parallel e(n) \parallel^2$. This forms the basis of the closed-phase LP analysis [11]. Once we get a_ks, we can get $e(n)$, which is an estimate of $u_g(n)$. Specifically, rewriting equation 2.2, we get

$$E(z) = S(z).[1 - \sum_{k=1}^{p} a_k z^{-k}] \tag{2.4}$$

Observe that, from equation 2.4, $e(n)$ can be viewed as the output of a filter with transfer function $1/H(z)$, when the speech signal $s(n)$ is the input. Thus, the method involves canceling the effect of the vocal tract filter by applying an inverse filter, and hence the term inverse filtering.

However, the closed-phase LP analysis method requires the knowledge of the closed-phase of the glottal cycle. If we have the electroglottograph (EGG) signal simultaneously recorded with the speech signal, we can get the instants of glottal closure, also called the glottal closure instants (GCIs) (the negative peaks of the differentiated EGG signal correspond to the GCIs). Using these, the closed-phase can be determined as the few samples after the GCI. Thus, we can estimate the VS by closed-phase LP analysis, as in [26, 20]. But, it is not practically feasible to have the EGG recorded simultaneously with speech in most cases. For this reason, we have to use a GCI estimation technique to estimate the closed-phase. Since GCI estimation algorithms have an error associated with them, this method suffers from occasional wrong estimates of the GCIs leading to wrong estimates of the closed-phase.

Also, some phonation types, e.g, the breathy phonation, do not involve complete closure of the vocal folds. Even in other phonation types, there might be incomplete closure, as observed for a particular case in Figure 2.1. Since this method relies on the presence of a clear and a long enough closed phase, this method fails in such cases [1]. Thus, to get an estimate of $u_g(n)$ which is independent of the nature of the closed-phase of the glottal cycle, in this study, we use a technique based on the voice source linear prediction model developed by Ananthapadmanabha [6], which is described in later sections.

2.2 Review of VS-based features for speaker recognition

Once the VS signal is estimated from the speech signal, one can extract relevant speaker-specific features from it to use in a speaker recognition task. Many general characterizations of the VS have been proposed. Among them, the time domain models of the VS pulse (glottal pulse) like the Liljencrants-Fant (LF) model [27] and the model proposed by Ananthapadmanabha [28] are noteworthy. The glottal pulse is also characterized in the frequency domain by defining many measures. Among them, the parabolic spectral parameter proposed by P. Alku [29] and the measures characterizing the spectral decay of the VS [20, 30] are of interest.

Characterizations like the ones mentioned above have been used to extract speaker-specific features in speaker recognition studies, as mentioned in section 1.3. A detailed review of VS-based speaker-specific features can be found in [31]. The time-domain samples of the LP residual themselves have been used as features in [22, 32]. However, the authors in [22, 32] have reported results on databases which were not available to us. Thus, in this thesis, we review three studies in the literature, whose authors have all reported results on the same databases (which were available) using the same classifier. We have chosen these three particular studies to compare the features proposed in these with the features proposed later in this thesis.

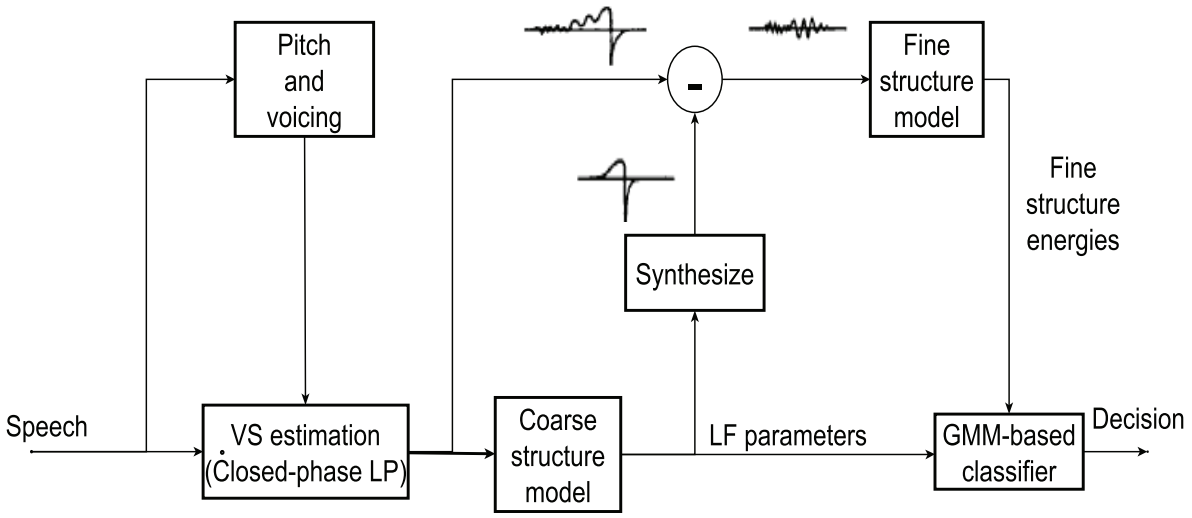


Figure 2.2: Block diagram of the feature extraction process developed by Plumpe et. al. (taken from [3])

2.2.1 LF model parameters

Plumpe et. al. [3] have performed detailed SID studies on the 168-speaker TIMIT test set (The entire TIMIT database consists of 630 speakers, but the authors report results on 168 speakers, and these results can be assumed to be scalable over the entire database). Figure 2.2 shows the block diagram of the method they have used to extract the speaker-specific features. The VS is obtained by closed-phase LP analysis as described in section 2.1.1. The VS (glottal pulse) waveform is decomposed into the 'coarse' and the 'fine' structures. The time domain LF model which they have used to estimate the coarse structure of the VS is described below.

The LF model is a piecewise continuous model of the VS pulse. A single glottal cycle shown in Figure 2.3 is modeled as follows:

$$\begin{aligned} v_{LF}(t) &= 0, && T_{c-1} \leq t < T_o \\ &= E_o.e^{\alpha(t-t_o)}.sin[\omega_o(t-T_o)], && T_o \leq t < T_e \\ &= -E_1[e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)}], && T_e \leq t < T_c \end{aligned} \tag{2.5}$$

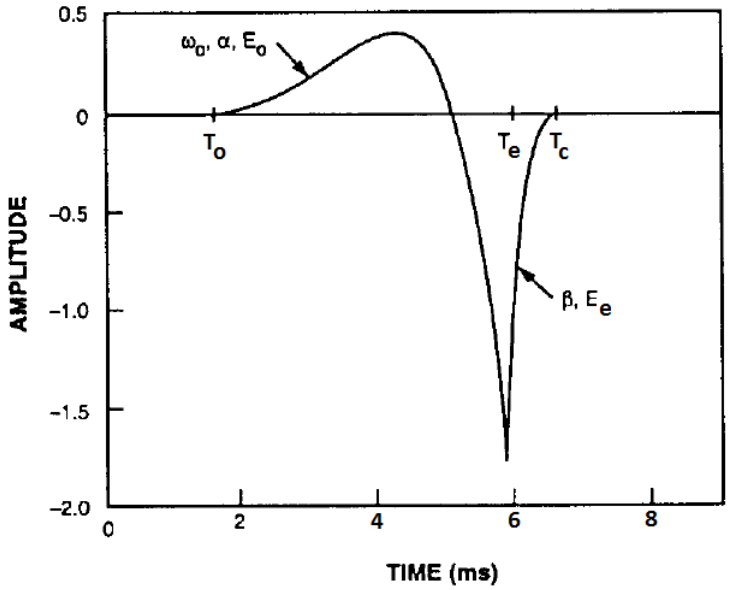


Figure 2.3: A single glottal cycle with the LF model parameters corresponding to different phases (taken from [3])

where the time instants T_o, T_e, T_c are the instants of glottal opening, negative peak and glottal closure, respectively, as shown in Figure 2.3. T_{c-1} represents the end of the return phase of the previous glottal cycle, and also the start of the current cycle. From equation 2.5, we can see that: (a) the phase from $T_{c-1} - T_o$, i.e, the closed phase is modeled simply as zero, (b) the phase from $T_o - T_e$, i.e, the open phase is modeled as an exponentially rising sinusoid with frequency ω_o and growth factor α and (c) the phase from $T_e - T_c$, i.e, the return phase is modeled as an exponential decay with a parameter β.

Considering each pitch period as a frame (pitch synchronous analysis), the LF model is fit over the estimated VS pulse by applying a non-linear least squares minimization technique called NL2SOL [33]. Thus, by model fitting, we get estimates of α, ω_o, β and E_e (negative peak amplitude of the glottal pulse). In addition to these, denoting the time period of the cycle by T, three parameters are defined: (a) the *close quotient* $(CQ) = (T_o - T_{c-1})/T$, (b) the *open quotient* $(OQ) = (T_e - T_o)/T$, and (c) the *return quotient* $(RQ) = (T_c - T_e)/T$. These represent the portion of time the vocal folds are closed, open and returning to the closed position in a glottal cycle. Thus, a total of 7 parameters are obtained from the coarse structure model.

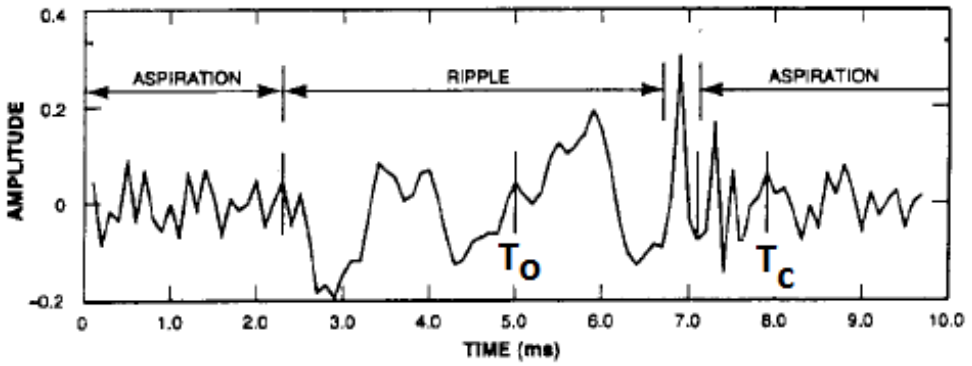


Figure 2.4: Fine structure of the VS showing aspiration and ripple (taken from [3])

Table 2.1: SID performance of LF model parameters on the TIMIT test set

Features used	**SID accuracy (%)**		
	Male	**Female**	**Average**
Coarse	58.3	68.2	63.3
Fine	39.5	41.8	40.7
Coarse+Fine	69.1	73.6	71.4

The coarse structure is synthesized using the estimated model parameters, and is subtracted from the estimated VS to get the fine structure of the VS. The fine structure is modeled as shown in Figure 2.4. The energies of the fine structure of the VS pulse over the closed, open and return phases, normalized w.r.t the total energy of the VS are considered as the fine structure features. These energies capture the aspiration and ripple energy present in the VS pulse.

With the coarse and fine structure features, each speaker is modeled by a Gaussian mixture model (GMM), and a maximum likelihood-based decision rule is used for classification [34] (this is explained in detail in subsequent chapters). The SID performance is shown separately for the male and female TIMIT subsets in Table 2.1.

From Table 2.1, we see that the coarse structure features alone give an average accuracy (over the male and female subsets) of 63.3%, from which it is evident that the shape of the VS contains an appreciable amount of speaker information. Also, the fine structure features alone give an average accuracy of 40.7%. The coarse and fine features taken together give an

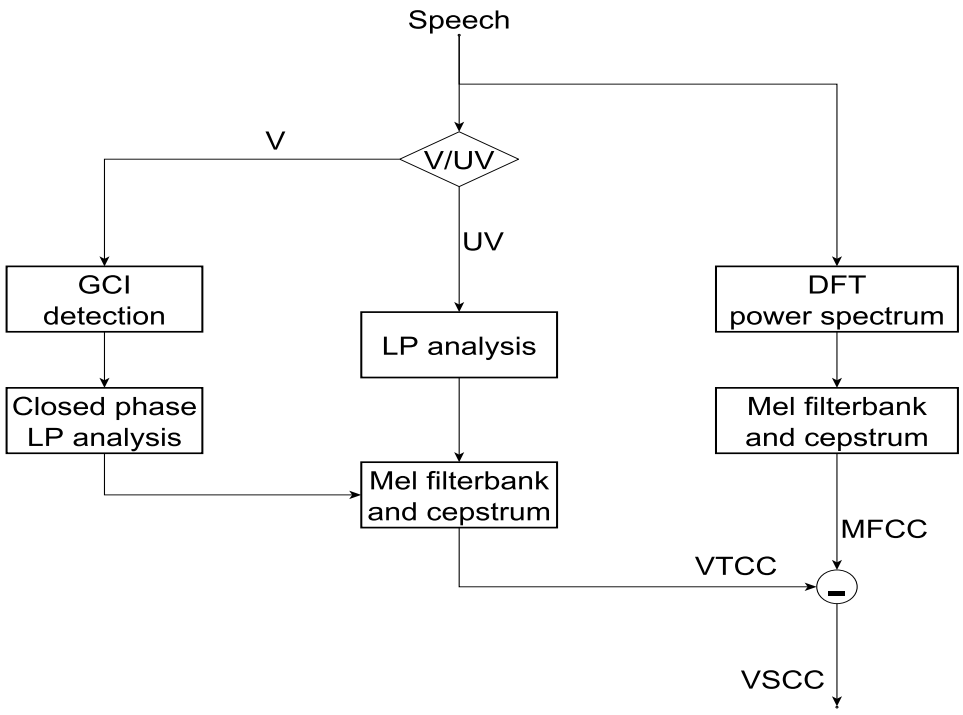


Figure 2.5: Block diagram of the method to extract VSCCs (taken from [4])

average accuracy of 71.4%, which is an improvement over that given by the coarse structure features by 8.1% (71.4%-63.3%). This shows that the fine structure also has significant speaker information, and is lost while fitting an idealized pulse shape using a time domain model.

2.2.2 Voice source cepstral coefficients

Gudnason et. al. [4] extracted VS-based cepstral features and used them in an SID task (using GMMs as speaker models, similar to [3]) over the TIMIT and the 138-speaker YOHO databases. In this section, only the results on the TIMIT test set are presented to compare with the LF model parameters described previously. A block diagram of the feature extraction process is shown in Figure 2.5.

With 32 ms frames and a 10 ms frame shift, MFCCs are extracted from the speech signal. A voiced/unvoiced decision algorithm is applied on the speech signal to separately process the voiced and unvoiced regions. On the voiced regions, the DYPSA algorithm [35] is applied

to estimate the GCIs, and these GCI estimates are used to determine the closed-phase. Closed-phase LP analysis is applied to estimate the speech spectral envelope, on which the mel-filterbank is applied. The cepstrum of the filterbank energies is taken to represent what they term the 'vocal tract cepstral coefficients' (VTCCs). This is because the LP analysis is applied when there is no input from the VS, and the speech spectral envelope corresponds to the contribution from the vocal tract alone. On the unvoiced regions, the spectral envelope is estimated by covariance LP analysis and its mel cepstral coefficients are extracted, which are the VTCCs. Since convolution in the time domain is addition in the cepstral domain, if we subtract the VTCCs from the MFCCs, we get a cepstral representation of the source, which they call the voice source cepstral coefficients (VSCCs).

The VSCCs are used as feature vectors, and, on the TIMIT database, they give an accuracy of 94.9%. Comparing this with the results given by the LF model parameters, clearly, the VSCCs capture more speaker information.

2.2.3 Deterministic plus stochastic model (DSM) of the LP residual

Drugman et. al. [5] have proposed a decomposition of the VS (impulse train view - usual LP residual) into the deterministic and the stochastic components, which they call the DSM. Figure 2.6 shows a block diagram of the process to extract the deterministic and stochastic components.

First, the GCI detection algorithm SEDREAMS [36] is applied on the speech signal, and the GCIs are used as pitch marks. The duration between two successive pitch marks is considered as a pitch period and pitch synchronous LP residual frames are extracted. These are pitch normalized using resampling and principal component analysis (PCA) is applied on the pitch normalized LP residual frames, and the eigenvector corresponding to the largest eigenvalue (termed by the authors as the eigenresidual) is called the deterministic component [1]. Parallelly, the pitch normalized LP residual frames are high-pass filtered, and the energy envelope of the high-pass filtered frames is considered the stochastic component.

[1]The eigenresidual is also used to synthesize speech in a HMM-based speech synthesis framework in [37]

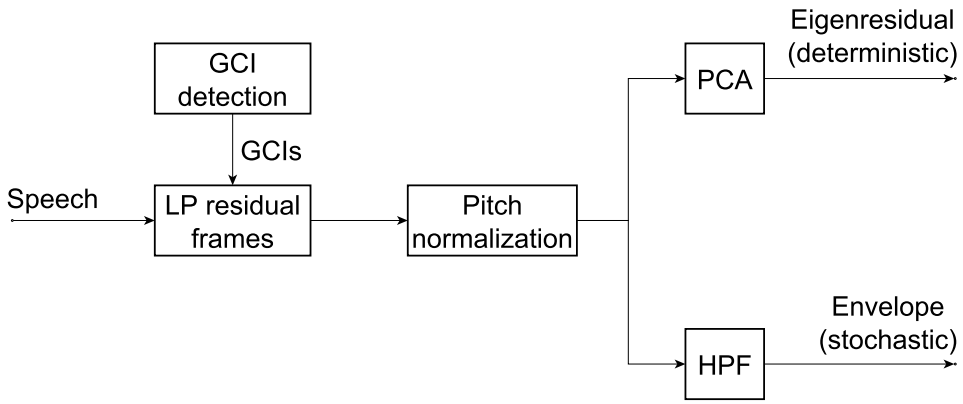


Figure 2.6: Block diagram of the method to extract the deterministic and stochastic components of the LP residual (taken from [5])

The deterministic and stochastic components are obtained for each speaker in the training phase, and are used as the speaker representations. In the test phase, speech from the unknown speaker is processed in the same way to get the deterministic and stochastic components of the LP residual. These are compared with each of the speaker representative deterministic and stochastic components using the normalized error energy as the distance metric D.

The distances between the deterministic and the stochastic components are fused as follows:

$$D(i) = \alpha.D_d(i) + (1 - \alpha).D_s(i) \tag{2.6}$$

where α is a scalar between 0 and 1, $D_d(i)$ is the distance between the deterministic components of the training speaker i and the test utterance, and $D_s(i)$ is the same between the stochastic components. The test utterance is declared to be from the speaker corresponding to the least $D(i)$.

The value of α is chosen by trial and error. It is varied from 0 to 1, and the value which gives the highest SID accuracy is taken to be the optimal α. For an optimal α of 0.25 on the TIMIT test set, the authors report an impressive SID accuracy of 98.0%.

Table 2.2: SID performance of parametric and non-parametric features

Features	**LF model parameters**	**VSCC**	**DSM**
SID accuracy (%)	71.4	94.9	98.0

2.2.4 Parametric vs non-parametric characterizations

Characterizations of the VS (or VS-based features) can be classified broadly into two categories: (a) parametric - meaning they are parameters of an idealized model fit over the time domain waveform or the spectrum of the VS, and (b) non-parametric - meaning they are not the parameters obtained by model fitting, but are characterizations derived purely from the estimated VS signal, usually by means of some kind of a transformation on the data.

Notice that, among the three features presented above, only the LF model parameters are parametric. The VSCCs are a cepstral domain transformation and are not extracted by model fitting, and hence are non-parametric. The deterministic and the stochastic components, though they are from the so-called deterministic plus stochastic 'model', are not parameters of a model fit, but are the eigenresidual waveform obtained by PCA (which is a dimensionality reducing transformation on the data), and an energy envelope, which is also a waveform. The performances of the three features are summarized for ease of reference in Table 2.2.

We can clearly see from Table 2.2 that the parametric features perform poorer than the non-parametric features. Particularly, since the LF model parameters and the VSCCs are derived from glottal pulses, the fact that the VSCCs perform better than the LF model parameters implies that the VS contains speaker information in the waveform as a whole, a significant amount of which is lost when the coarse and fine structures of the VS are considered separately by fitting a time domain model to it.

But notice that the VSCCs are an indirect characterization of the VS, and the DSM uses the LP residual frames (impulse train view of the VS). Thus, we seek a new non-parametric direct characterization of the glottal pulses, since the glottal pulses are a closer representation of the actual physiology involved in the speech production process.

Chapter 3

Characterization of the voice source by the DCT

In order to characterize the VS, we first need an estimate of the VS. The closed-phase LP analysis method suffers from the dependency on the speech signal, as mentioned in section 2.1.1, and we seek a method which avoids such issues. Thus, at this juncture, we revisit the source-filter model.

3.1 Revisiting the source-filter model - the voice source linear prediction model

Recall the two views of the VS, shown in Figure 1.2. Let us take a closer look at the 'glottal pulse train' view of the VS and the source-filter model associated with it. As explained in section 1.1.2, with a log scale for the magnitude, the VS spectrum has a spectral tilt, and most of the spectral tilt in the speech signal spectrum is due to the VS. Using glottal area measurements, the spectral tilt/rate of fall of the glottal volume velocity has been found to be -12 $dB/octave$, and the calculated spectral envelopes have been presented in [8]. However, as mentioned in section 1.1.2, we consider the glottal volume velocity derivative as the VS. The differentiation operation introduces a positive spectral tilt of 6 $dB/octave$, and hence the glottal volume velocity derivative has an overall spectral tilt of -6 $dB/octave$.

The reason for considering the volume velocity derivative as the VS[1] is explained below.

The speech signal is radiated out from the lips as a pressure wave, and it is this pressure variation that we record from the microphone. Approximately, the transfer function from the volume velocity through the lips $u_l(t)$ to the pressure $p(t)$ at a distance D from the lips, in the spectral domain, is given by the following equation [8]:

$$P(\omega) = U_l(\omega)\frac{\rho\omega}{4\pi D}K_T(\omega) \tag{3.1}$$

where $U_l(\omega)$ and $P(\omega)$ are respectively the Fourier transforms of $u_l(t)$ and $p(t)$, ρ is the density of air and $K_T(\omega)$ is a correction factor. Neglecting the effect of $K_T(\omega)$, we can say that, at a given D, $P(\omega)\ \alpha\ \omega U_l(\omega)$, which in the time domain translates to $p(t)\ \alpha\ \frac{du_l(t)}{dt}$. Thus, in the source-filter model equation (equation 1.1), if $s(n)$ is the recorded speech pressure waveform, $u_g(n)$ is the glottal volume velocity derivative.

With this relationship established between the pressure and the volume velocity, the VS can be estimated from the speech signal. To this end, a model called the voice source linear prediction (VSLP) model has been proposed in [6]. In the VSLP framework, the synthesis model is the same as the 'glottal pulses view' of the VS shown in Figure 1.2. The analysis model is slightly different from usual LP analysis, and this difference is explained in the next subsection.

3.1.1 The integrated linear prediction residual

The analysis model in the VSLP framework is shown in Figure 3.1. In the case of usual LP-based inverse filtering, we pre-emphasise the speech signal, obtain LP coefficients and use them to inverse filter the pre-emphasized speech signal. But, in VSLP, we pre-emphasize the speech signal, obtain LP coefficients and use them to inverse filter the non-pre-emphasized speech signal itself.

Figure 3.1 shows the VSLP analysis model, along with the spectra of the speech signal (Figures 3.1(b) and (d)) and the magnitude response of the vocal tract filter (Figure 3.1(c)),

[1]This insightful explanation was given by Dr. T.V. Ananthapadmanabha orally during a personal interaction

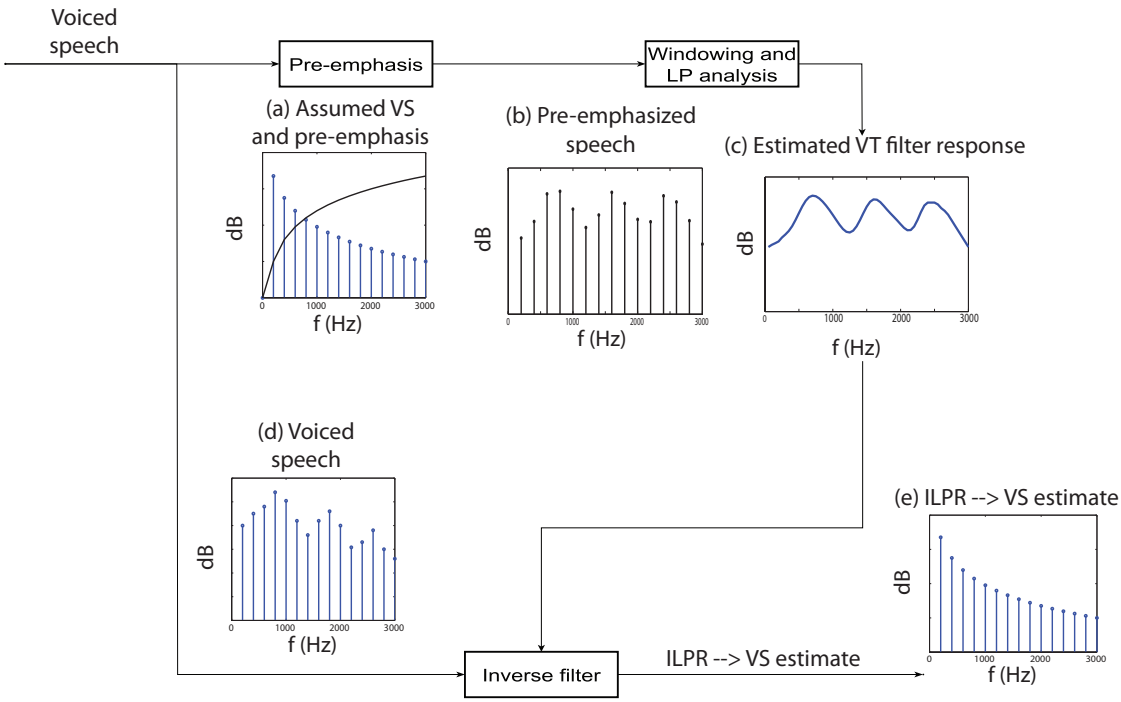


Figure 3.1: VSLP analysis model along with spectra at each stage - the ILPR is an estimate of the VS (taken from [6])

at every stage in the analysis. The difference between usual LP and VSLP can be seen in Figure 3.2. When the speech signal is pre-emphasized, the spectral tilt contributed by the VS is nullified. Thus, the VS component of the pre-emphasized speech signal has a flat harmonic spectrum, corresponding to the 'impulse train view' of the VS. Hence, when LP coefficients estimated from the pre-emphasized speech signal are used to inverse filter the pre-emphasized speech signal, as in the case of usual LP analysis, the LP residual we obtain is similar to an impulse train. Also, note that the spectral envelope of the pre-emphasized speech signal corresponds to the vocal tract filter's magnitude response in the 'glottal pulses view' of the VS, and the LP coefficients obtained represent the vocal tract filter. Thus, when these LP coefficients are used to inverse filter the non-pre-emphasized speech signal, as in the case of VSLP, we cancel out the effect of the vocal tract filter to obtain a harmonic spectrum with the spectral tilt retained, thus giving an estimate of the 'glottal pulses'. From the relationship between pressure and volume velocity, we know that, if the speech signal is the recorded pressure waveform, this estimate corresponds to the volume velocity derivative,

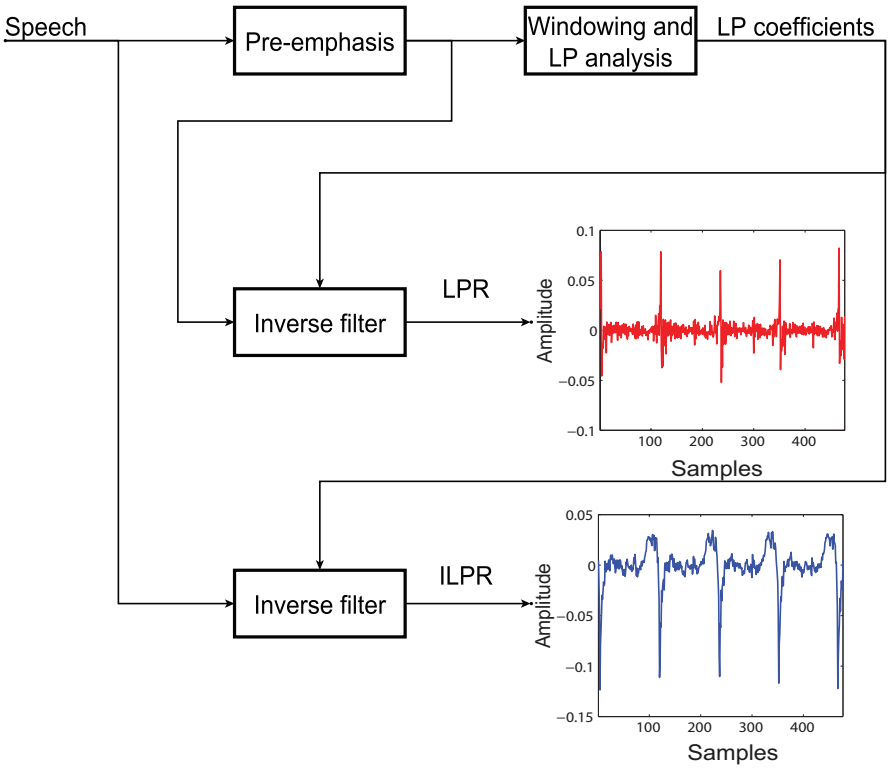


Figure 3.2: Difference between usual LP analysis and VSLP analysis - The LPR and the ILPR

which is the desired VS.

Observing the spectra in Figure 3.1, we see that the speech signal spectrum has a tilt of ≈ -6 $dB/octave$, a majority of which is contributed by the VS. After pre-emphasis, the spectral tilt in both the spectra of the speech signal and the VS is nullified, and this leads to the LP residual to have a flat spectrum. But, when we inverse filter the non-pre-emphasized speech signal itself, the spectral tilt of ≈ -6 $dB/octave$ is preserved in the LP residual, leading to an estimate of the VS (Figure 3.1(e)). Thus, the difference between the spectra of the usual LP residual and the LP residual obtained from the VSLP framework is that the latter has a ≈ -6 $dB/octave$ spectral tilt, while the spectral tilt of the former is almost zero. For this reason, the latter can be viewed as a low-pass or an integrated version of the former, and hence the name integrated linear prediction residual (ILPR).

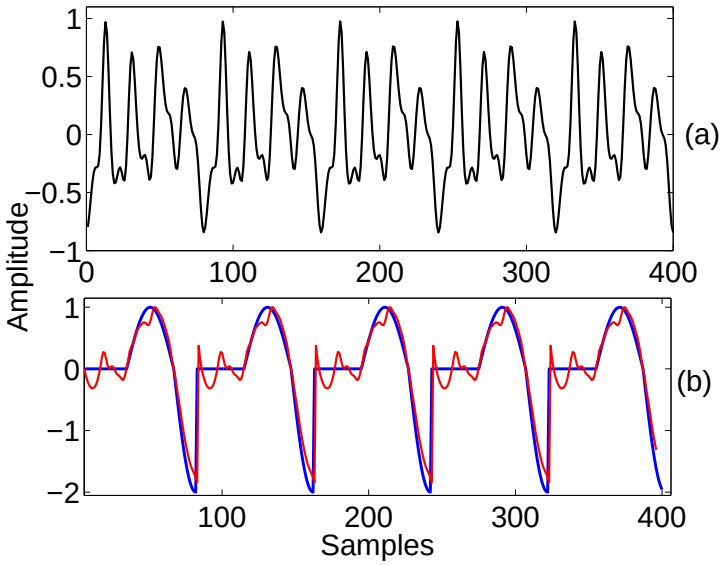


Figure 3.3: ILPR of synthetic speech. (a) A synthesized vowel segment. (b) The VS pulse train used for synthesizing the vowel (blue) and the estimated ILPR (red).

3.1.1.1 An example of ILPR on synthetic speech

A VS pulse train is synthesized using the VS pulse model proposed in [28], with the pitch frequency chosen to be 200 Hz, and the CQ, OQ and RQ chosen to be 0.25, 0.6 and 0.15, respectively. A vocal tract filter is synthesized using a resonator cascade. The formant frequencies are chosen to be F1 = 850 Hz, F2 = 1500 Hz, and F3 = 2813 Hz, corresponding to the vowel /a/. The formant frequencies are taken to be the center frequencies of the resonators, and the bandwidths are chosen to be 80 Hz, 80 Hz, and 120 Hz, respectively.

The synthesized vowel is shown in Figure 3.3(a), along with the VS pulse train used for synthesis in Figure 3.3(b). The synthesized speech signal is pre-emphasized with a simple first difference filter, i.e, a filter with impulse response $p_e(n) = [1 - 1]$. Taking the negative peak locations of the VS pulses to be the pitch marks, three consecutive pitch periods of the pre-emphasized speech signal are used to estimate the LP coefficients, using a Hanning window for autocorrelation LP analysis. The non-pre-emphasized speech signal is inverse filtered using these coefficients, and only the middle period is retained, because the first p samples of the inverse filtered (where p is the LP order) signal are erroneous, and the Hanning window tapers for the first and third periods but the middle period is not too much

affected by the windowing.

It is also worth noting that the pre-emphasis filter $p_(n)$, having an anti-symmetric and finite impulse response, has a linear phase response. Thus, the pre-emphasis operation just introduces a finite delay, without affecting the phase of the signal.

Figure 3.3(b) also shows the estimated ILPR overlaid on the VS pulse train used for synthesis. We can observe that, except for a small deviation near the positive peak in the opening phase, the open phase is estimated almost perfectly, whereas there are some ripples in the closed phase. These ripples and deviations are mainly due to improper cancellation of the first formant. Otherwise, we can see that the ILPR is a fairly good estimate of the VS.

In this study, due to its ease of estimation and elegance, and also due to the fact that it is unaffected by the nature speech signal unlike the LP residual obtained using closed-phase analysis, we use the ILPR as the estimate of the VS.

3.2 The DCT of ILPR: A characterization of the VS

From previous SID studies, we know that there is speaker information in the VS waveform as a whole. While concluding the previous chapter, we mentioned that we are seeking a non-parametric direct characterization of the VS in order to extract this speaker information. In this study, we explore the DCT as a means to characterize the VS.

3.2.1 Motivation for using the DCT to characterize the VS

The DCT is a real frequency domain transformation of a signal. It has been demonstrated to be successful when used as a feature extractor in many recognition problems [38, 39]. It has many impressive properties:

1. Energy compaction - It captures most of the signal energy in a very few coefficients due to its excellent energy compaction property [40]. Thus, it enables us to represent the VS signal in a few coefficients.

2. Invertibility - The DCT is invertible and hence captures the shape, phase and amplitude of the signal. Since we know that the shape of the VS contains speaker information,

the DCT may be able to capture it.

3. Fast implementation - There are many efficient and fast algorithms to compute the DCT [40], and hence computational complexity is not an issue.

In addition to these properties, the following hypothesis is another motivation for using the DCT to characterize the VS: Timbre is defined by ANSI[2] as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar." The voice of a speaker is analogous to the timbre of a musical instrument. It has been shown in several studies [41, 42, 43] that the levels of different 'harmonics' can be used to characterize a musical instrument's timbre. This so called 'harmonic series' is used for recognizing musical instruments in solos and even in the presence of accompaniments. We hypothesize here that a speaker's voice may be determined by the size, shape and position of the vocal folds, and the vocal fold vibrations reflect the effect of these factors. Since the VS represents the vocal fold vibrations, the relative levels of the harmonics of the VS may determine the speaker's identity. Since the basis vectors of the DCT are harmonically related cosines, the magnitudes of the DCT coefficients can be regarded as the amounts of different harmonics present in the VS. The signs of the DCT coefficients capture the phase or the 'shape' information, which is known to contain speaker information. Moreover, the DCT coefficients are not the parameters of any model, and are thus non-parametric. Thus, the DCT coefficients are an attractive characterization of the VS, especially with speaker recognition as the target application.

3.2.2 Pitch synchronous analysis

The DCT is known to be sensitive to shifts in start and end points, which means that the DCT coefficients change even for small left or right shifts in the boundaries of the signal. Due to this issue, if we frame the signal with a constant frame size, say 20 ms (which is generally used for LP analysis), we encounter the following problem.

Figure 3.4(a) shows a 20 ms segment of the ILPR estimated from the synthetic vowel shown in Figure 3.3, and Figure 3.4(b) shows the same segment cyclically shifted by half a

[2]American National Standards Institute

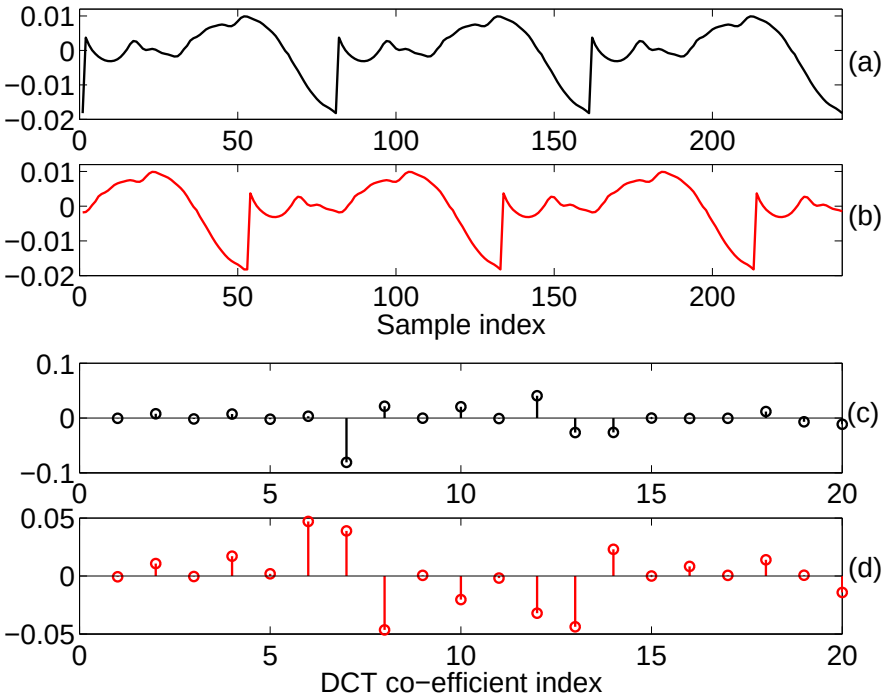


Figure 3.4: (a) ILPR estimated from a 20 ms segment of a synthetic vowel and (b) its cyclically shifted version; (c) and (d): The first 20 DCT coefficients of the signals in (a) and (b), respectively.

pitch period. Figures 3.4(c) and (d) show the DCTs of the ILPRs in Figures 3.4(a) and (b), respectively. We can easily see that the DCT coefficients in Figures 3.4(c) and (d) are very different from each other. But, even though the signal is circularly shifted, the VS pulse shape has not changed, and we want the DCT-based characterization to be the same in both the cases. This is not possible if we use a constant frame size analysis, and hence we use a pitch synchronous analysis, i.e, we estimate pitch marks and take two successive pitch marks to be the start and end points of frames, and obtain the DCT coefficients of each such frame. Since each frame corresponds to a pitch period, the analysis is termed pitch synchronous analysis.

Also, pitch synchronous analysis allows an unambiguous interpretation of the DCT coefficients as harmonics, which is not possible in the case of a constant frame size analysis. This interpretation is explained in Appendix A. Further, pitch synchronous DCT has been successfully used for pitch modification in [44].

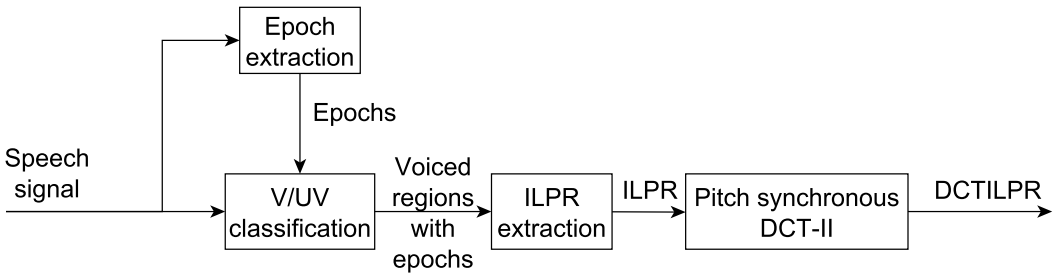


Figure 3.5: Block diagram of the method to extract the pitch synchronous DCT of ILPR

3.2.3 Obtaining the pitch synchronous DCT of ILPR

Figure 3.5 shows the block diagram of the method used to extract the pitch synchronous DCT of ILPR. First, an epoch extraction algorithm [45] is applied on the speech signal to obtain the epoch locations. Then, a voiced/unvoiced (V/UV) classification algorithm is applied as in [46] and the voiced regions are obtained. Since the epoch locations in the voiced regions correspond to GCIs, they are used as pitch marks for pitch synchronous analysis. Considering the interval between two successive GCIs as a pitch period, the ILPR is obtained with three consecutive pitch periods used for LP analysis, inverse filtering the three pitch periods and retaining only the middle period of the inverse filter output. A Hanning window is used for LP analysis, and a first difference filter is used for pre-emphasis. The analysis region is then shifted one pitch period to the right, and the same LP analysis and inverse filtering process is carried out till all the voiced regions are traversed. Finally, the DCT-II is applied on the ILPR pitch synchronously, after normalizing each period of the ILPR by its positive peak value. Thus, after the feature extraction process, we get as many DCT vectors as the pitch periods.

The epoch extraction and the V/UV classification algorithms used here are described briefly in the remainder of this section.

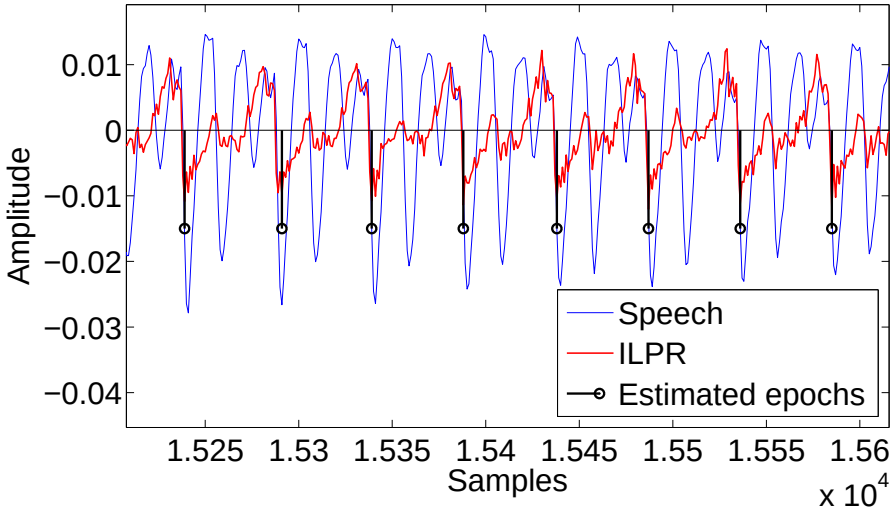


Figure 3.6: A TIMIT speech segment with the ILPR and the estimated epoch locations; the epochs can be seen at the negative peaks of the ILPR

3.2.4 Pre-processing used to obtain the DCT of ILPR

3.2.4.1 Epoch extraction algorithm

The state-of-the-art dynamic plosion index (DPI) algorithm proposed by Prathosh et. al. [45] is used for epoch extraction. The plosion index (PI) is a dimensionless ratio used to detect transients or impulse-like signals, and is used to detect plosion burst instants in [45]. The DPI is a sequence of values of the PI as a function of the number of samples chosen to compute the PI. The ILPRs extracted from 20 ms speech signal frames is half-wave rectified to retain only the negative going parts, since the epochs correspond to the negative peaks of the ILPR. An arbitrary instant is chosen as the initial epoch location. The DPI is then computed on the half-wave rectified ILPR, and, as shown in [45], the instant of the valley of the DPI associated with the maximum peak-valley difference corresponds to the epoch location. The process is repeated till the end of the speech signal is reached.

The epoch location estimates obtained by the DPI algorithm are shown overlaid on a voiced segment of the speech signal in Figure 3.6. Also shown for reference is the estimated ILPR. We can observe that the negative peak locations of the ILPR are estimated as the epoch locations.

Note that we estimate the ILPR on 20 ms frames in this process, which is observed to be sufficient for epoch extraction, since only the negative peaks of the ILPR are of interest. However, we need to refine the ILPR extraction process if we want to use the ILPR as an estimate of the VS, and hence we adopt the procedure of taking three pitch periods for LP analysis and retaining only the middle period of the inverse filter output, as described in section 3.1.1.1.

It is worth mentioning here that this algorithm gives epoch estimates in both voiced and unvoiced regions, but only the epoch estimates in the voiced regions are meaningful. Thus, we need to apply a V/UV classification algorithm, which is described next.

3.2.4.2 Voiced/Unvoiced classification algorithm

The V/UV algorithm used here is based on the maximum normalized cross correlation (MNCC) of successive inter-epoch intervals, and has been used successfully in [46]. It is based on the observation that the cross-correlation between successive inter-epoch intervals is high in voiced regions and low in unvoiced regions. Since we have the epoch location estimates from the epoch extraction algorithm, we can compute the cross-correlation between two successive inter-epoch intervals. This cross-correlation is normalized with respect to the energy in the first inter-epoch interval, to get the normalized cross-correlation (NCC) at different lags. Since there may be slight phase differences between the two successive inter-epoch intervals due to minor errors in epoch estimation, the NCC need not be maximum at zero lag, and thus we take the maximum of the NCC across all the lags as the similarity measure, and hence the name MNCC.

Figure 3.7 illustrates a TIMIT speech segment and its phonetic annotation, and also shows the evolution of MNCC over successive inter-epoch intervals. It is evident that the MNCC is above a threshold (0.6 in this case) for voiced phones and below it for unvoiced phones, which shows the effectiveness of the MNCC-based V/UV classification.

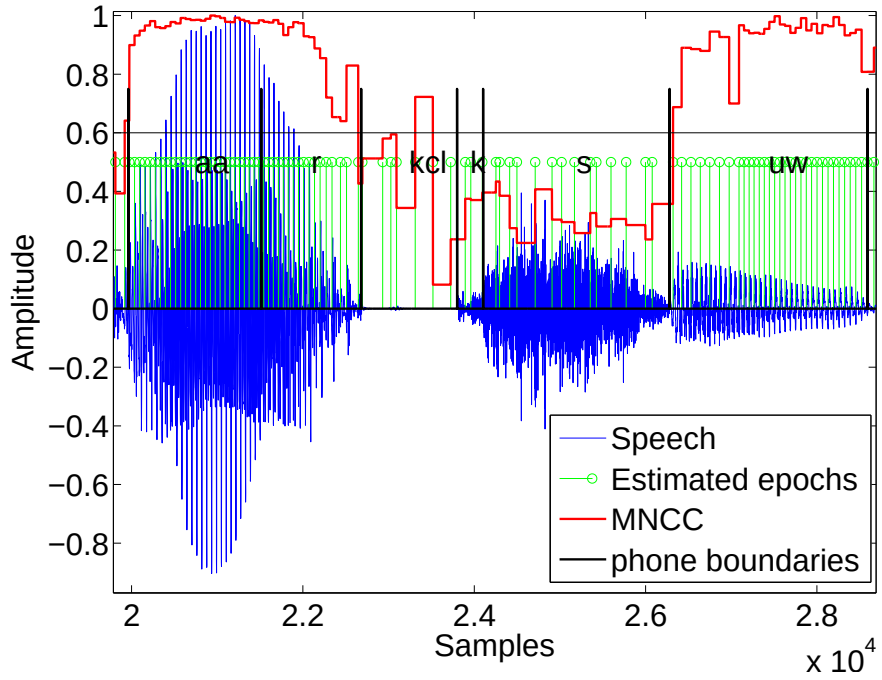


Figure 3.7: A segment from a TIMIT utterance and the corresponding MNCC plot overlaid on it; observe that MNCC>0.6 for voiced phones and <0.6 for unvoiced phones.

3.3 The DCT of ILPR as a feature vector for SID

It is worthwhile to visually observe how the DCT coefficients of the ILPR change from speaker to speaker. Since we know that there is speaker-specific information in the VS, characterizing it by the DCT must reflect the same.

3.3.1 Variability of the ILPR

Figure 3.8 shows three ILPR segments taken from three different utterances of a TIMIT speaker, and their corresponding pitch synchronous DCT coefficients, obtained from the second period of the shown waveforms. It is observed that the ILPRs are similar and so are the DCT coefficients. Figure 3.9 shows the ILPR segments of three different TIMIT speakers, and their corresponding pitch synchronous DCT coefficients (of the second period). Here, we can observe that the ILPR segments are different and so are the DCT coefficients, e.g, the ILPR of speaker 1 (red) has a low amplitude negative peak, and its corresponding DCT

coefficients have a lesser amplitude than those of speakers 2 (black) and 3 (blue). The DCT coefficients of speaker 3 show a positive-negative alternating pattern in the higher coefficients, which is not shown by those of speaker 2. Thus, from Figures 3.8 and 3.9, we can say that the DCT coefficients capture distinct information for different speakers.

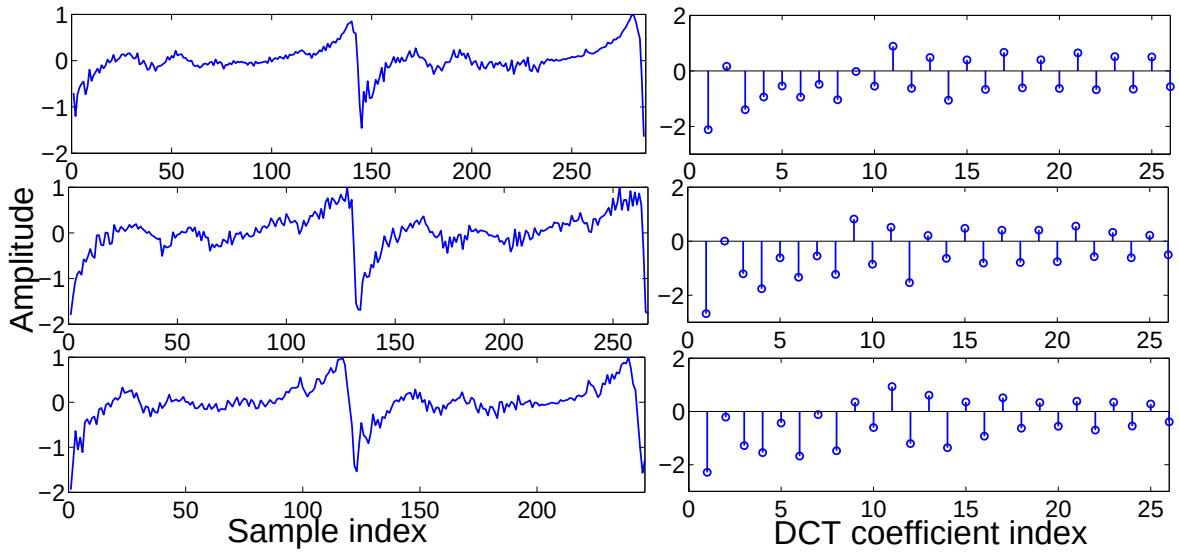


Figure 3.8: ILPR and its DCT from different phonetic contexts in voiced segments from three different utterances of a single TIMIT speaker. The shapes of the ILPRs are similar and so are the DCT coefficient vectors.

However, the ILPR does show some intra-speaker variability too. Although the ILPR for a single speaker shows a similar shape as shown in Figure 3.8 in a majority of the regions, it does vary in some regions. This may be because of two major factors:

1. Since we are not estimating each formant and cancelling them explicitly in the inverse filtering process, there may be traces of uncancelled F1 and F2 overriding on the actual VS pulse, which lead to differences in ILPR for different phones.

2. The VS pulse shape itself may change for the same speaker due to the speaker possessing different phonation types, as is shown in [47].

The differences in the ILPR for the same speaker are reflected in the DCT coefficients, and hence the DCT coefficients exhibit intra-speaker variability. Figure 3.10 shows a 2-D

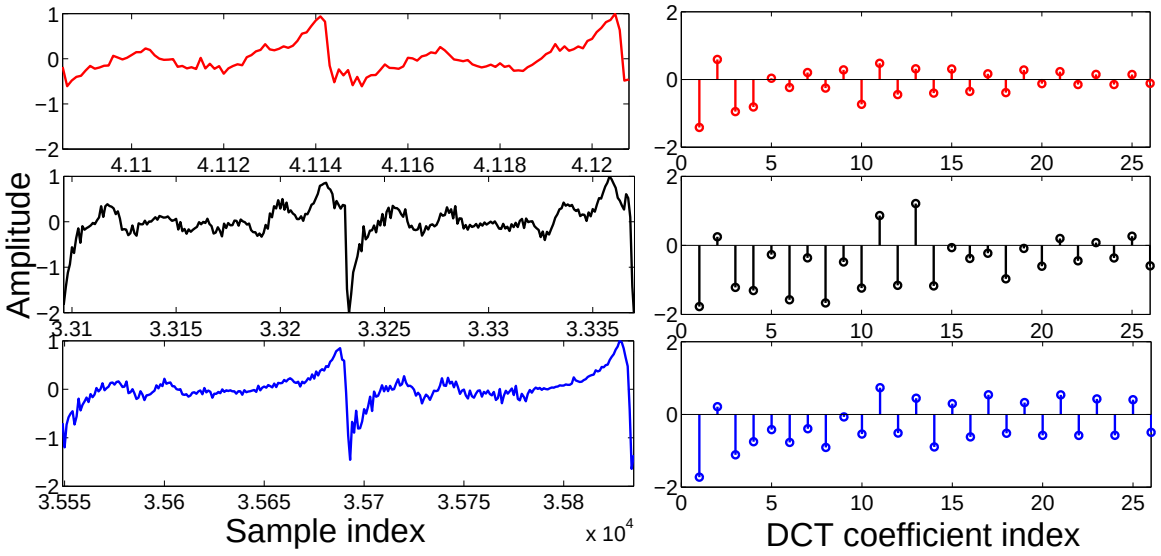


Figure 3.9: ILPR and its DCT from the vowel /a/ in the word 'wash' for three TIMIT speakers. The ILPR shape varies from speaker to speaker and the DCT coefficients vary with waveform shape.

scatter plot of the DCT coefficients 5 vs 10 of the ILPR for a single TIMIT speaker. We can see that there are multiple clusters in the feature space (marked by ellipses). Figure 3.11 shows a 3-D scatter plot of the DCT coefficients 1 vs 5 vs 10 of the same speaker. In both figures, the cluster marked by the red and black ellipses have a significant data concentration, while the cluster marked by the green ellipse is loosely formed. The multiple clusters are seen more clearly in Figure 3.11. This illustrates the intra-speaker variability of the DCT coefficients.

However, there are differences in the data distribution between different speakers. Figure 3.12 shows the 3-D scatter plot of DCT coefficients 1 vs 5 vs 10 for a different TIMIT speaker from the one shown in Figure 3.11. We can see that there is only a single prominent cluster in Figure 3.12, and almost all the data points have negative value for DCT coefficient 10, whereas from Figure 3.11, we can see that there are multiple clusters and the DCT coefficient 10 has positive as well as negative values for the previous speaker. Thus, if we are able to capture these differences in the data distribution by a speaker modeling technique, we may be able to use the DCT of ILPR directly as a feature vector for SID.

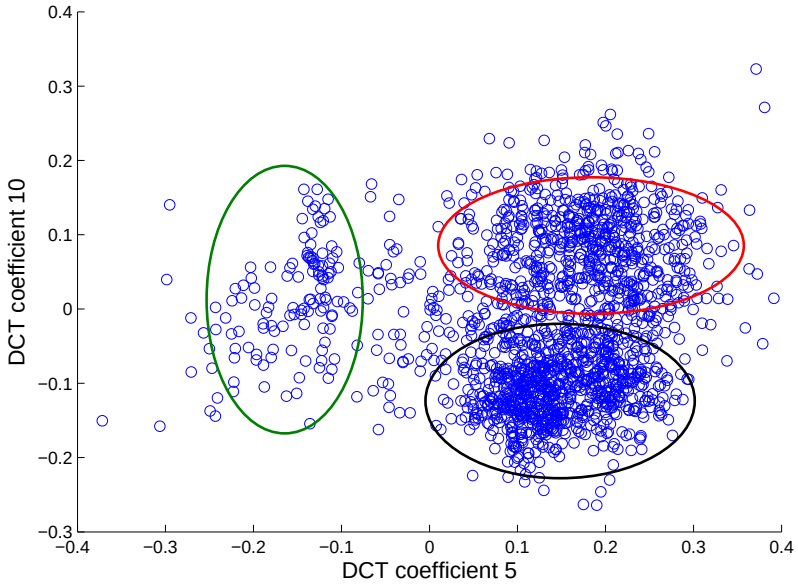


Figure 3.10: 2-D scatter plot of DCT coefficients 5 vs 10 of the ILPR cycles from a TIMIT speaker. Cluster boundaries are roughly indicated by ellipses.

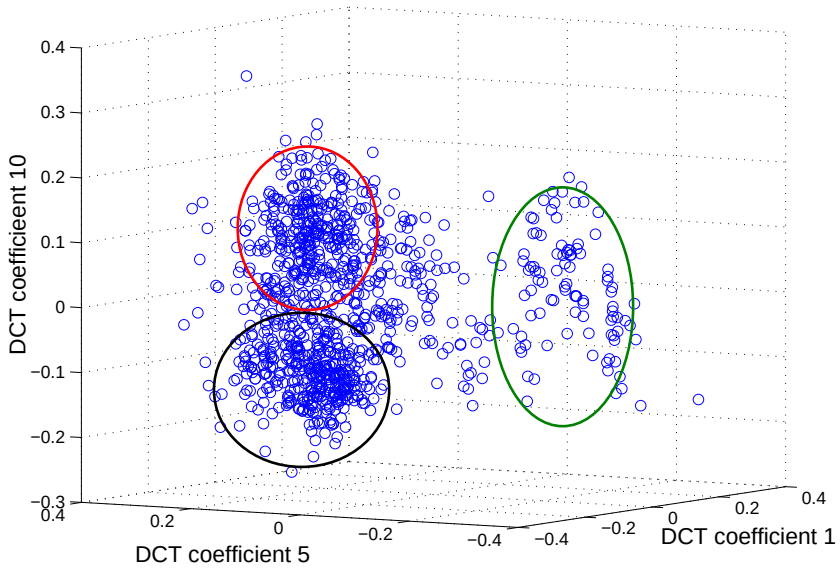


Figure 3.11: 3-D scatter plot of DCT coefficients 1 vs 5 vs 10 of the ILPR cycles from a TIMIT speaker. Cluster boundaries are roughly indicated by ellipses.

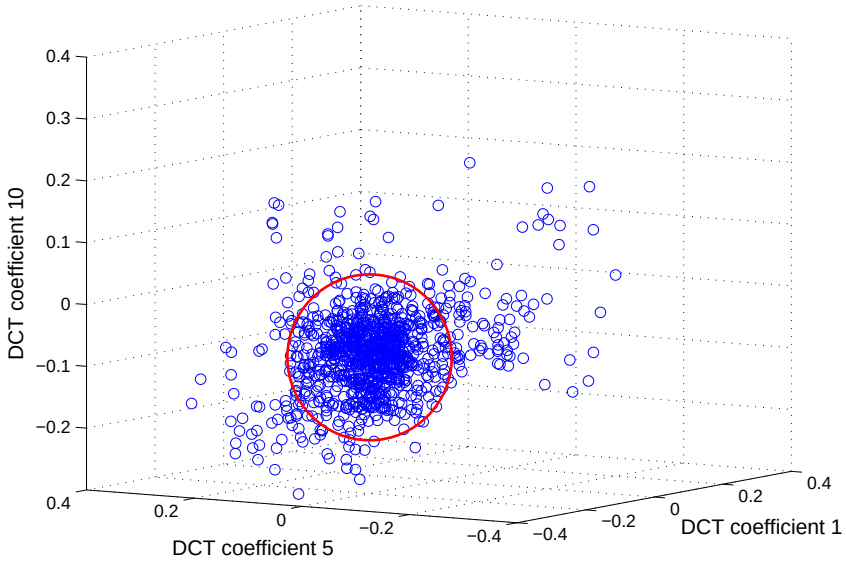


Figure 3.12: 3-D scatter plot of DCT coefficients 1 vs 5 vs 10 of the ILPR cycles from another TIMIT speaker. There is only one prominent cluster and the data distribution is different from that of the previous speaker.

3.3.2 Number of DCT coefficients

One of the motivations to characterize the VS by the DCT is the energy compaction property of the DCT, by which we can represent the signal by a few coefficients. Thus, we need not consider all the DCT coefficients as the feature vector. In addition, since we would be using them in a pattern recognition framework, it is ill-advised to consider a very high dimensional feature vector (with $f_s = 16\ kHz$ and a pitch frequency $F_0 = 200\ Hz$, we will have 80 samples in a pitch period, and since we take the N-point DCT, we will have an 80 dimensional feature vector). Thus, we need to determine the number of DCT coefficients sufficient for a 'good' feature vector.

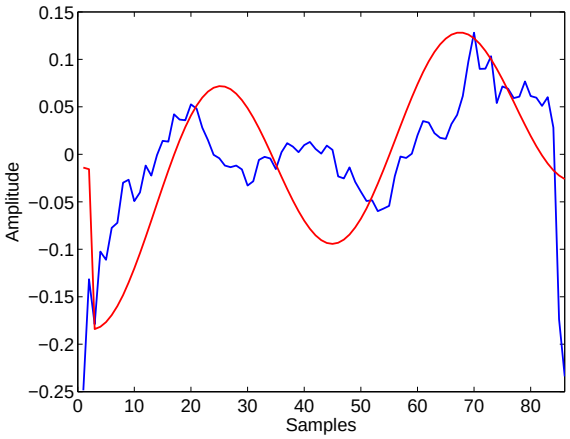
Since the DCT is an invertible transform, we can reconstruct the ILPR by its DCT coefficients using the inverse DCT. In order to estimate the number of DCT coefficients for a good feature vector, we reconstruct the ILPR using the inverse DCT by varying the number of DCT coefficients, i.e, we consider the first M DCT coefficients, zero out the remaining $N - M$, and take the N-point inverse DCT to get the reconstructed ILPR. Let us visually

observe how the reconstruction changes with M.

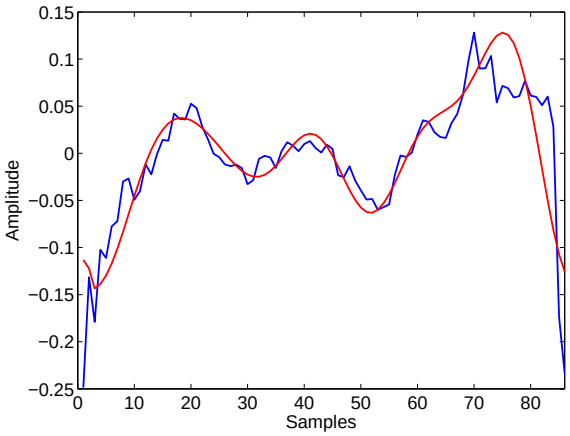
Figure 3.11 shows the reconstructed ILPR overlaid on the original for various values of M from 4 to 50 in different steps (which corresponds to reconstruction using 2 harmonics to 25 harmonics of the fundamental - see Appendix A). We can observe that, in general, the reconstruction gets closer to the original as M increases, which is because we capture more and more of the finer details (higher frequency components) as we increase M. However, from Figure 3.11(c), the gross ILPR pulse shape seems to be well captured with $M = 16$ (within 8 harmonics), which gives us a hint that $M = 16$ may be a good choice for considering the DCT coefficients as a feature vector (note that, by 'gross shape', we do not mean the 'coarse structure' in the sense of Plumpe et. al. [3] as explained in section 2.2.1, but rather the smooth shape of the ILPR, excluding the very fine variations).

Further, the mean percentage of signal energy captured in the reconstruction was computed using the pitch synchronous DCT coefficients of ILPRs from all the voiced frames of 100 TIMIT speakers. It is shown in Figure 3.12, for M varying from 5 to 40. We see that it saturates after $M = 15$, with the reconstruction capturing more than 90% of the signal energy for $M \geq 15$, adding more compelling evidence to consider the first 16 coefficients as the feature vector. However, since SID is the target application, we need to determine how many coefficients contain 'sufficient speaker information', which is an entirely different issue.

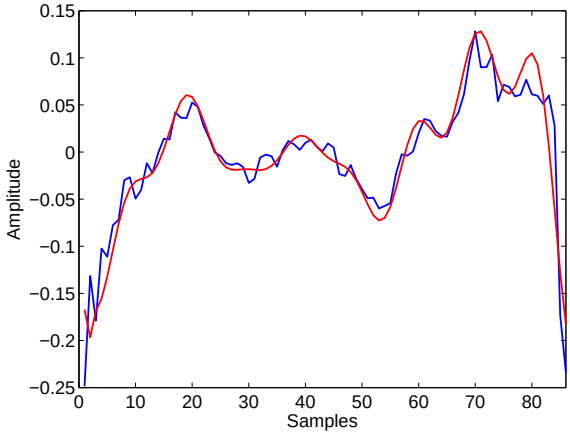
Thus, to be able to use the DCT of the ILPR as a feature vector for SID, two issues are to be addressed: (a) the optimal M has to be determined, and (b) a modeling technique which can capture the feature distribution has to be decided. These are resolved and the results of SID experiments using the DCT of ILPR as the feature vector are presented in the next chapter.



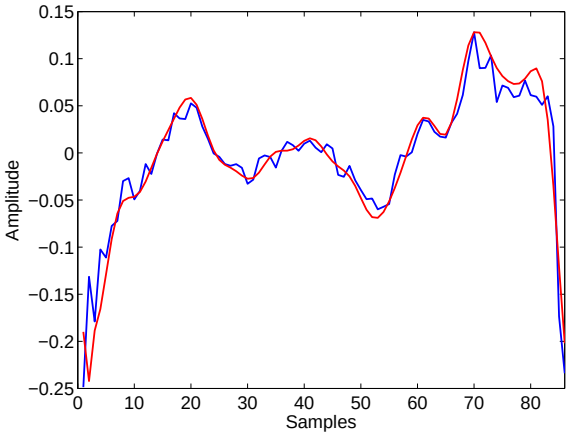
(a) ILPR (blue) and its reconstruction (red) with the first 4 DCT coefficients



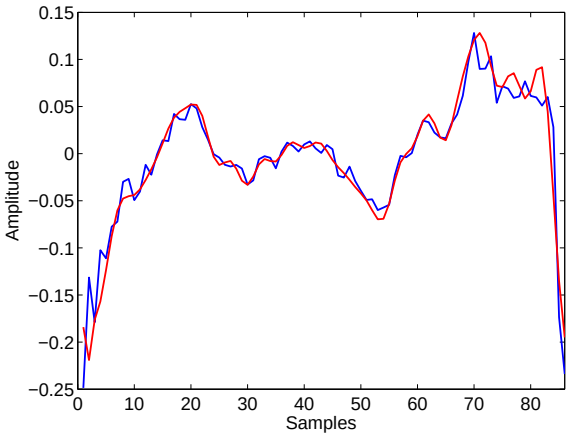
(b) ILPR (blue) and its reconstruction (red) with the first 10 DCT coefficients



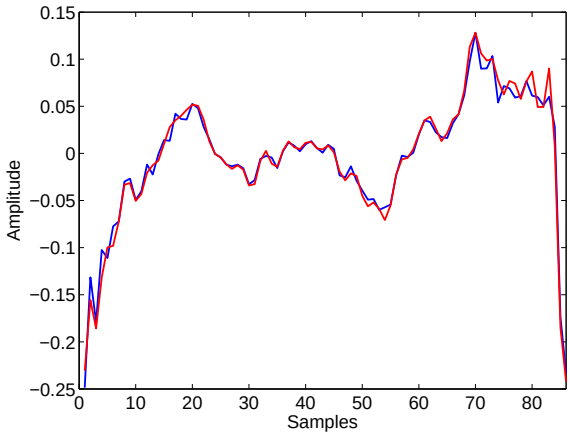
(c) ILPR (blue) and its reconstruction (red) with the first 16 DCT coefficients



(d) ILPR (blue) and its reconstruction (red) with the first 24 DCT coefficients



(e) ILPR (blue) and its reconstruction (red) with the first 34 DCT coefficients



(f) ILPR (blue) and its reconstruction (red) with the first 50 DCT coefficients

Figure 3.11: One period of ILPR and its reconstruction from its truncated DCT, varying the number of DCT coefficients retained.

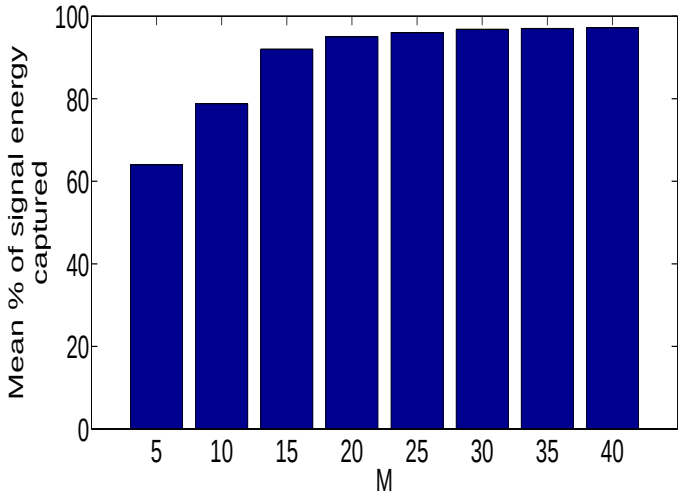


Figure 3.12: Mean of the percentage of signal energy captured as a function of M, computed using data from 100 TIMIT speakers.

Chapter 4

Experiments and results

4.1 Gaussian mixture models for SID

Visual observation of scatter plots shown in Figures 3.10, 3.11 and 3.12 indicates that the distribution of the DCT coefficients of ILPR may distinguish speakers. To capture the distribution, a natural choice would be a Gaussian mixture model (GMM), since any arbitrary distribution in an M-dimensional space can be approximated by a mixture of M-dimensional Gaussians with different weights, means and covariance matrices [48]. Thus, we use GMMs to model each speaker, as explained in [34], reproduced here for convenience.

Let Z be the number of speakers and λ_i the GMM learnt from the M-dimensional training data of the i^{th} speaker. Thus,

$$p(\vec{x}|\lambda_i) = \sum_{j=1}^{G} w_{j,i}.b_{j,i} \tag{4.1}$$

where $b_{j,i}(\vec{x}) = \frac{1}{(2\Pi)^{M/2}|\Sigma_{j,i}|^{1/2}} exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_{j,i})'{\Sigma_{j,i}}^{-1}(\vec{x} - \vec{\mu}_{j,i})\}$ is the j^{th} Gaussian mixture component of λ_i with mean $\vec{\mu}_{j,i}$ and covariance matrix $\Sigma_{j,i}$, $w_{j,i}$ is the weight associated with $b_{j,i}$, G is the number of mixture components of λ_i and $p(\vec{x}|\lambda_i)$ represents λ_i evaluated at $\vec{x}$. The parameters $\{\vec{\mu}_{j,i}, \Sigma_{j,i}, w_{j,i}\}$ of each λ_i are learnt using the expectation maximization (EM) algorithm [49]. Thus, after the training phase, we will have a GMM for each speaker.

In the testing phase, the test utterance $\vec{X}$ is presented to each λ_i, and the likelihood $p(\lambda_i|\vec{X})$ is computed as follows:

By Bayes' rule, we have

$$p(\lambda_i|\vec{X}) = \frac{p(\vec{X}|\lambda_i)p(\lambda_i)}{p(\vec{X})} \tag{4.2}$$

Assuming the test utterance can belong to any speaker with equal probability, and since we present the same test utterance to all the GMMs, $p(\lambda_i)$ and $p(\vec{X})$ are constant $\forall\ i$. Thus, equation 4.2 reduces to $p(\lambda_i|\vec{X})\ \alpha\ p(\vec{X}|\lambda_i)$.

Now, if we assume that the test feature vectors are all statistically independent, we have $p(\vec{X}|\lambda_i) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda_i)$, where $\vec{x}_t$s are the individual test feature vectors and T is the total number of test feature vectors. Taking logarithm, we have,

$$log[p(\vec{X}|\lambda_i)] = \sum_{t=1}^{T} log[p(\vec{x}_t|\lambda_i)] \tag{4.3}$$

Thus, $log[p(\lambda_i|\vec{X})]\ \alpha\ \sum_{t=1}^{T} log[p(\vec{x}_t|\lambda_i)]$, where the RHS is the log-likelihood, which involves terms λ_i evaluated over $\vec{x}_t$, and can be easily calculated. Let L_i denote the log-likelihood of speaker i. Thus, if S is the decision,

$$\begin{aligned} S &= \text{argmax}_i\, log[p(\lambda_i|\vec{X})] \\ \implies S &= \text{argmax}_i\, L_i \end{aligned} \tag{4.4}$$

From equation 4.4, a logical decision rule to use would be to classify the test data as belonging to the speaker whose model gives the maximum likelihood. Figure 4.1 illustrates the test phase.

4.2 Issues with GMMs - The covariance matrix type and the number of Gaussians

In any GMM implementation, the type of the covariance matrix used (full or diagonal) and the number of Gaussians G have to be chosen. We conducted initial experiments with full covariance matrices by determining the optimal G using cross-validation, as explained below.

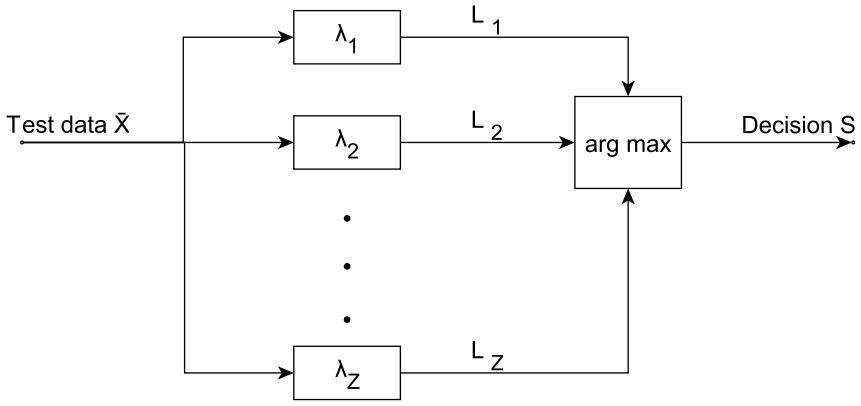


Figure 4.1: Test phase of the SID system using GMM speaker models

All the experiments were performed on the TIMIT test set, which is described in the next section. The first 24 DCT coefficients (excluding the 0^{th}) were chosen as the feature vector (the reason for this will be explained later).

4.2.1 Full covariance matrices

Using full covariance matrices, when a fixed G was chosen for all the speakers, it was observed that the SID accuracy was poor (around 60%). This is because different speakers have different number of clusters in the feature space, as shown by the feature distribution plots in Figures 3.10, 3.11 and 3.12. Since we do not know the number of clusters beforehand, and also since G is not exactly proportional to the number of clusters (one Gaussian may be sufficient for some clusters, but others may require more number of Gaussians to capture the variability), we employ a cross-validation strategy and a maximum likelihood-based decision to determine the optimal G for each speaker.

Figure 4.2 shows the use of maximum likelihood to determine the optimal G for a given training set. Utterances 5-10 are selected for training, and 1 and 2 for the test. Utterances 3 and 4 are used as validation data, and are presented to GMMs learnt using the training data with G varying over a fixed range $(A - B)$ in Figure 4.2). The GMM which gives maximum likelihood is chosen to be the one with the optimal G, which best represents the speaker. To determine the range $A - B$, we performed the following experiments.

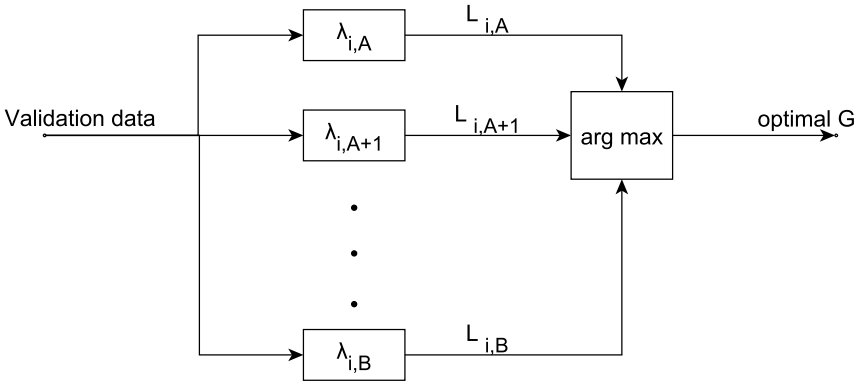


Figure 4.2: Cross-validation strategy to determine the optimal number of Gaussians for modeling each speaker. $\lambda_{i,j}$ represents the GMM with j Gaussians for speaker i.

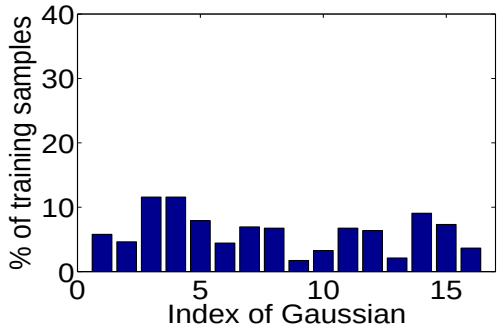
4.2.1.1 Histograms of number of training samples assigned to different Gaussians

The training data of two TIMIT speakers were used to learn the respective GMMs, varying G. As before, $b_1, b_2, ..., b_G$ are the G Gaussian components of the GMM. A training sample $\vec{x}$ was assigned to one of the Gaussians based on the following rule:

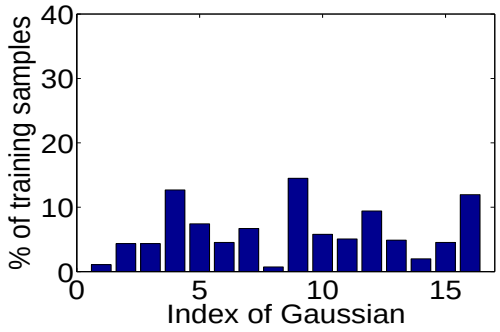
1. Compute the posterior probability of each Gaussian given the sample observation, i.e, $p(b_i|\vec{x})$, $i = 1, 2, ..., G$ using Bayes' rule.

2. Assign the sample $\vec{x}$ to the Gaussian having the maximum posterior probability.

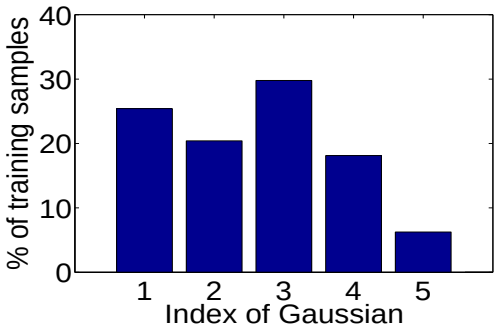
The histograms of the number of training samples assigned to different Gaussians, obtained with $G = 16$ are shown in Figures 4.3(a) and (b). It is apparent that only 5-10 Gaussians are assigned more than 10% of the training samples, for both speakers. Figures 4.3(c) and (d) show the histograms with $G = 5$ for speaker 1 and $G = 9$ for speaker 2, and we observe that, even with $G = 5$, there are only 4 Gaussians which are assigned more than 10% of the training samples. This demonstrates that the intra-speaker variability of the DCTILPR is very less, and we may not need more than 10 Gaussians to capture the data distribution of any speaker.



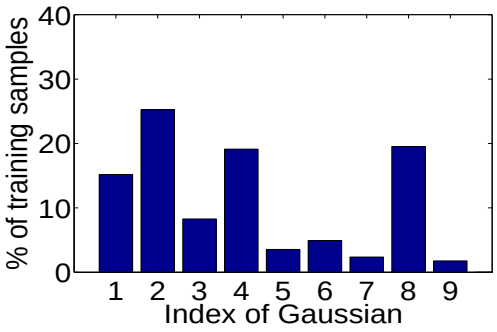
(a) Histogram of training samples from speaker 1 assigned to various Gaussians with $G = 16$



(b) Histogram of training samples from speaker 2 assigned to various Gaussians with $G = 16$



(c) Histogram of training samples from speaker 1 assigned to various Gaussians with $G = 5$



(d) Histogram of training samples from speaker 2 assigned to various Gaussians with $G = 9$

Figure 4.3: Histograms of training samples from two TIMIT speakers assigned to various Gaussian components of the GMMs for different G.

4.2.1.2 SID experiments with optimal G determined in different ranges

Motivated by the observations from the histogram plots, we performed SID experiments on different number of speakers in the TIMIT test set. For different ranges of G, the optimal G was determined using cross-validation, as shown in Figure 4.2. From Table 4.1, we see that the SID accuracy decreases as the number of speakers increases, since more speakers cause more confusions. This is a general trend observed in all the SID experiments conducted. Observe that the range 10 to 20 gives better performance than 26 to 35, and the range 5 to 10 gives the best performance. This confirms the limited intra-speaker variability of the DCTILPR. The SID accuracy for the entire 168-speaker database is 86.6%.

Table 4.1: SID performance on the TIMIT test set using GMMs with full covariance matrices, with the optimum G determined in different ranges using cross-validation. The range 5-10 gives the best result.

Range of G	Number of speakers	SID accuracy (%)
26-35	16	91.0
	25	75.0
	49	68.5
	62	65.6
10-20	16	95.0
	25	87.0
	49	88.9
	62	85.2
5-10	30	**100**
	62	**93.6**
	100	**90.0**
	168	**86.6**

4.2.2 Diagonal covariance matrices

Using diagonal covariance matrices, we assigned a fixed G for all the speakers and varied G over a range, and the SID accuracy versus G is shown in Figure 4.4. It is apparent that the accuracy increases with increasing G, attains a maximum at $G = 16$, and saturates beyond $G = 16$. Also, we found that the performance did not change even if we determined G by using the cross-validation strategy similar to the one used for full covariance matrices, varying G in the range 5-20. This may be because the diagonal covariance matrices do not capture the cross-correlations between the features, and thus the effect of overfitting the data may be lesser than that when full covariance matrices are used. In other words, if the number of Gaussians chosen is greater than the optimal value, an error in the estimation of cross-correlations in the covariance matrix of even even one of the Gaussians may lead to a large overfit of the data in the full covariance case, but the diagonal covariance matrices are immune to this. It is expected that GMMs with diagonal covariance matrices require more number of Gaussians to perform on par/better than those with full covariance matrices, because we need multiple diagonal covariance matrices to fit the data distribution captured

by one full covariance matrix [48].

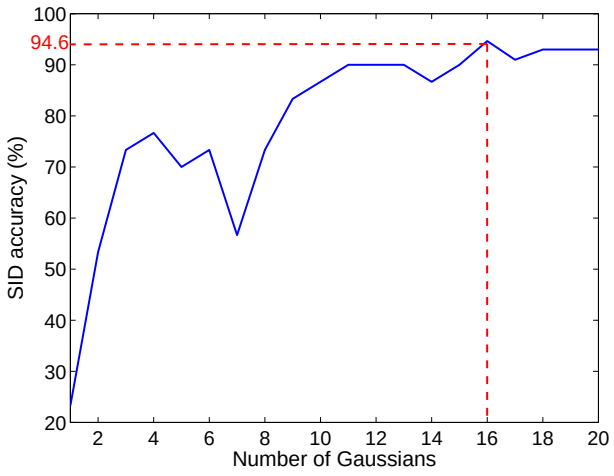


Figure 4.4: SID accuracy on the 168-speaker TIMIT test set versus the number of Gaussians (chosen to be the same for all speakers) with diagonal covariance matrices.

The SID accuracy for $G = 16$ is **94.6%** on the entire 168-speaker set, which is much higher than that given by the full covariance matrices (86.6%), even with optimal G determined in the latter case. This can be the result of two factors:

1. The DCT coefficients, which are the features here, are naturally decorrelated. This is because the DCT basis is similar to the Karhunen-Louve transform (KLT) basis, as shown in [40], and the KLT is known to be a statistically optimal decorrelating transform [50]. With full covariance matrices, we may be trying to capture cross-correlations which do not exist, and thus end up in estimating them wrongly.

2. Since we have to estimate cross-correlations also if we use full covariance matrices, the number of parameters of the GMM increases as the square of M, and there may be insufficient data to estimate all these parameters, but the data may be sufficient to estimate only the variances of individual features in the diagonal covariance matrices.

Thus, we decide that diagonal covariance matrices and $G = 16$ are to be used in further experiments. The result that the diagonal covariance matrices perform better than full covariance matrices is consistent with similar empirical observations made in [34, 4].

4.3 Description of SID experiments

SID experiments are conducted on three standard databases, and the performance of the DCT coefficients of the ILPR (henceforth referred as DCTILPR) is compared with that of other VS-based features. For the reasons explained in the previous section, GMMs with $G = 16$ are used as speaker models. As mentioned in section 2.2, the authors in [3, 4] have used GMMs as speaker models, and so it is fair to use GMMs to evaluate the performance of DCTILPR, which is another reason to use GMMs in this study. Before presenting the results, a brief description of the databases and the training and test data used in this study is given below.

4.3.1 Databases, training and test data

Three standard databases namely the TIMIT, YOHO and the NIST 2003 database are used.

4.3.1.1 The TIMIT database

The TIMIT database [51] is a well-known database used for speech and speaker recognition studies. It is designed to be phonetically rich, and has read speech recorded at 16 kHz from 630 speakers with 10 utterances from each speaker, each utterance being roughly 3 s in duration. However, the test subset of the database with 168-speaker data is chosen, since the authors who have proposed the VS-based features discussed in section 2.2 have presented their results on this subset, and we can compare the DCTILPR features with the existing VS-based features on the same dataset.

Utterances 3-10 from each speaker are taken as the training data and utterances 1-2 are presented as the test data to each speaker model learnt. Thus, the test data is just $(1/4)^{th}$ of the training data, which we may expect when the system is deployed in a practical scenario.

4.3.1.2 The YOHO database

The YOHO database [52] is a 138-speaker database, which is recorded for the purpose of speaker recognition studies. It is more challenging than TIMIT, because it has data from four different recording sessions, and there may be a change in the speaker's voice itself from

session to session, which is called session variability in the literature. Thus, there could be a mismatch between the same speaker's training data taken from one session and the test data taken from another session, and the results may be poorer than when both are taken from the same session, as we do in TIMIT. The speech consists of combination-lock phrases recorded at 8 kHz.

Again, the data is divided roughly into 80% for training and 20% for testing. There are 24 utterances per speaker per session, each roughly 2.5 s long. Utterances 1-20 are used for training and 21-24 for testing.

4.3.1.3 The NIST 2003 database

NIST (National Institute of Standards and Technology) conducts a speaker recognition evaluation (SRE) to assess the speaker recognition performance of different algorithms on a common platform. The data released with the NIST SRE 2003 [53] (sampled at 8 kHz) is used here to evaluate the performance on telephone quality noisy speech, since it contains excerpts of one side of a cell phone conversation, with each speaker speaking in different cell phone handsets. The features may change from one handset to the other, which is called handset variability, which may lead to mismatch between the training and the test data from the same speaker. There are 356 speakers, and around 110 s of training data and 60 s of test data per speaker.

4.4 The number of DCT coefficients for a 'good' feature vector

We performed the SID experiment on the TIMIT and YOHO databases and observed the accuracies by varying the number of DCT coefficients (M) considered for the feature vector. The training data was taken from session 1 and the test data was taken from session 2 in the YOHO database. On TIMIT, we can observe from Figure 4.5 that the accuracy reaches 92.8% for $M = 12$ and reaches a maximum of 94.6% at $M = 24$, which means that 98.1% of the maximum performance is achieved for $M = 12$. This indicates that most of

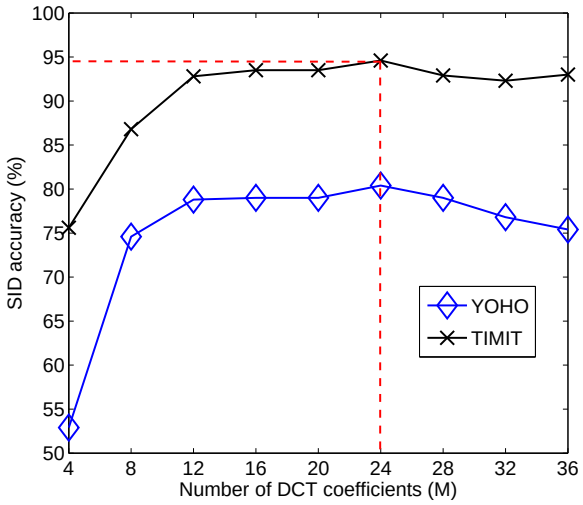


Figure 4.5: SID accuracy versus the number of DCT coefficients (M) on the 168-speaker TIMIT test set and the 138-speaker YOHO database. GMMs with 16 Gaussians having diagonal covariance matrices are employed as the speaker models.

the speaker information is contained in the first 12 DCT coefficients (6 harmonics), which can be explained to be a direct consequence of the low-pass nature of the VS. From Figure 3.11.(c) and the discussion in section 3.3.2, we saw that most of the gross shape of the VS is captured for $M = 16$, and combining this with the previous result we can say that most of the speaker information is contained in the gross shape of the VS. But, the surprising result is the performance drop for $M > 24$, which is counter-intuitive since we tend to think that the higher coefficients capture the finer structure of the VS and may account for more speaker information. This trend indicates that the very fine structure of the VS (captured by the harmonics greater than the 12^{th}) does not contain speaker information, but may cause more confusions between speakers instead.

Results on YOHO also show similar trends, but we observe that the SID accuracies are much lesser than what were observed for TIMIT (94.6% and 80.4% being the maximum accuracies on TIMIT and YOHO, respectively). This is because of the speaker variability from session 1 to session 2 of the YOHO database, showing that the DCTILPR features have significant session variability.

Table 4.2: Comparison of SID performance of different VS-based features on the 168-speaker TIMIT test set. GMMs with 16 Gaussians having diagonal covariance matrices are employed as the speaker models.

Feature	#misidentifications	Identification rate
TDVS	48	71.4%
VSCC	9	94.6%
DCTILPR	9	94.6%
DSM	**3**	**98.0%**

4.5 Results and discussion

4.5.1 Results on TIMIT

Table 4.2 shows the performance of the DCTILPR features compared with those of the time domain voice source (TDVS) features described in section 2.2.1, the VSCC features described in section 2.2.2 and the DSM-based features described in section 2.2.3. It is clear that DSM is the best performing feature with an SID accuracy of 98% (only 3 misidentifications). The DCTILPR performs equally well as the VSCC features with an accuracy of 94.6% (9 misidentifications), and much better than the TDVS features, thus showing more convincingly that the non-parametric features perform better than parametric features.

4.5.2 Results on YOHO

In [5], the DSM-based features have been shown to outperform the other features on the YOHO database. The performance of the DCTILPR features is compared with that of the DSM features in Table 4.3. If the model giving the maximum likelihood (or score) corresponds to the true speaker, the true speaker is labelled as a speaker in I position. Similarly, if the model giving the next higher likelihood (or score) corresponds to the true speaker, the true speaker is labelled as a speaker in II position, and so on. It is clear from Table 4.3 that, as the duration between the train and test sessions increases, the performance

decreases, showing the session variability of the features. We can also observe that more and more speakers get into the II and III positions as the gap between the sessions increases. But note that the DCTILPR gives a better performance than the DSM in all the cases, from which we can conclude that it has less session variability than the DSM features.

Table 4.3: Percentage of speakers classified in different positions with DCTILPR (with DSM) when test data is taken from different recording sessions in the 138-speaker YOHO database. GMMs with 16 Gaussians having diagonal covariance matrices are employed as the speaker models.

Session details	In I position	In II position	In III position
Same session	**100%** (99.7%)	0% (0.3%)	0% (0%)
1 session later	**80.4%** (69.3%)	3.6% (7.9%)	2.9% (5.2%)
2 sessions later	**73.9%** (64.3%)	2.9% (8.8%)	5.1% (4.6%)
3 sessions later	**72.5%** (58.7%)	5.1% (11.8%)	3.7% (4.4%)

The SID accuracies of the DCTILPR, DSM and VSCC features are averaged across the 4 sessions, and these average accuracies are presented in Table 4.4 for comparison. Results on TIMIT showed that the DSM performs best, whereas the DCTILPR performs best on YOHO. This difference may be because of two reasons:

1. In TIMIT, there is very little training data (24 s), which may be insufficient for speaker modeling with DCTILPR features. In YOHO, there is almost twice that amount of data (50 s), which perhaps leads to better speaker models being learnt, and hence the DCTILPR performs better on YOHO than on TIMIT. In fact, we observe that the DSM *also* shows a performance improvement from 98% on TIMIT to 99.7% on YOHO (same session test data).

2. There is less phonetic variability in YOHO than in TIMIT, because the speech is comprised of combination-lock phrases, e.g., nineteen-fiftyfive-fortytwo, which are only numbers, whereas TIMIT speech is designed to be phonetically rich. The ILPR may have more phonetic variability than the LP residual used to obtain DSM features, and hence DSM performs better than DCTILPR on TIMIT.

Table 4.4: SID accuracies of the DCTILPR, DSM and VSCC features on the 138-speaker YOHO database, averaged across the 4 sessions. GMMs with 16 Gaussians having diagonal covariance matrices are employed as the speaker models.

Feature	Average accuracy(%)
DCTILPR	**81.7**
DSM	73.0
VSCC	63.7

4.5.3 Results on NIST 2003

On the entire 356-speaker database, the DCTILPR gives a very low SID accuracy of 16%. We suspected this to be the effect of handset variability, and to test this we conducted the SID experiment on a 110-speaker subset of the database, with speakers having training data from one handset and test data from the same and different handsets. Table 4.5 gives the results on this subset. It can be seen that the accuracy drops from 71.8% when the test data is taken from the handset same as that of the training data (same handset condition) to 18.2% when the test data is taken from a handset different from that of the training data (different handset condition). This huge drop in accuracy clearly shows that the DCTILPR features suffer greatly from handset variability. The authors who have proposed other VS-based features have not reported results on NIST 2003.

Table 4.5: Comparison of SID performance of DCTILPR with MFCC features and their combination on the 110-speaker subset of the NIST 2003 database under the same and different handset conditions.

Condition	DCTILPR	MFCC	Classifier fusion $[\alpha.L_C + (1-\alpha).L_M]$
Same handset	71.8%	72.7%	Max.accuracy = **84.5**% $\alpha_{opt} = 0.67$
Different handset	18.2%	40.0%	Max.accuracy = 40.9% $\alpha_{opt} = 0.15$

4.5.3.1 Combination with MFCCs

To compare the performance of DCTILPR with vocal tract-based features, we consider the MFCCs of speech, which are perhaps the most widely used features in speech processing. The MFCCs are computed on 20 ms frames with 10 ms overlap, which is customary. But, they are obtained only from voiced regions to facilitate a fair and effective comparison with the DCTILPR, since the DCTILPR is obtained only from voiced regions and the MFCCs would get an undue advantage over the DCTILPR if they are obtained from unvoiced regions also. 13-dimensional MFCCs along with their delta and delta-delta coefficients are considered together as a 39-dimensional feature vector. From Table 4.5, we see that the MFCCs perform comparably to the DCTILPR in the same handset condition, whereas they perform much better than the DCTILPR in the different handset condition. Nevertheless, the drop in SID accuracy from 72.7% to 40% clearly shows that the MFCCs too suffer from handset variability, but not as much as the DCTILPR. This is mostly because the MFCCs capture only the magnitude, while the DCTILPR captures both the magnitude and phase information, and thus phase response variations between different microphones can also alter the DCTILPR significantly.

The classifier trained using DCTILPR is combined with the classifier trained using the MFCCs at the score level as indicated below:

$$L_C = \alpha L_D + (1 - \alpha) L_M \tag{4.5}$$

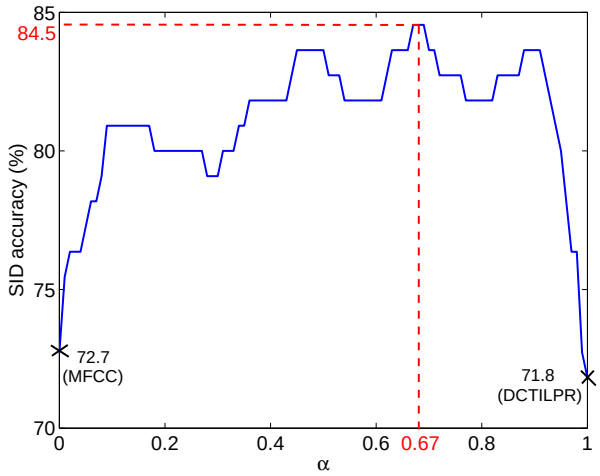


Figure 4.6: SID performance of the combination of DCTILPR and MFCC-trained classifiers as a function of α for the same handset condition on the 110-speaker subset of the NIST 2003 database.

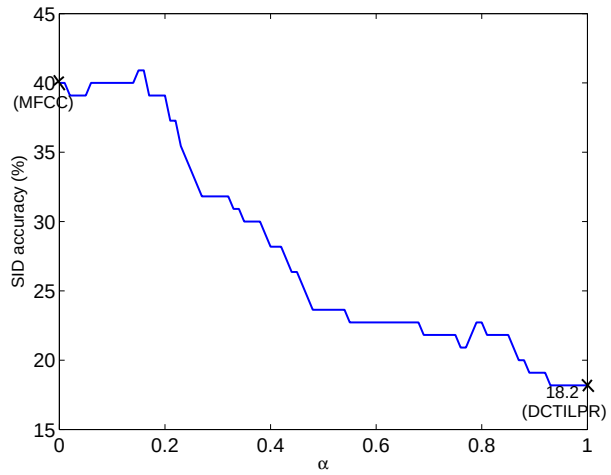


Figure 4.7: SID performance of the combination of DCTILPR and MFCC-trained classifiers as a function of α for the different handset condition on the 110-speaker subset of the NIST 2003 database.

where L_D and L_M are the likelihoods of the DCTILPR and MFCC-trained GMMs, and

the combined likelihood L_C is obtained as a linear convex combination of the two. α is a scalar varying from 0 to 1, and it represents the weightage given to the DCTILPR features, with the MFCCs having $(1-\alpha)$ as the weightage. Figure 4.6 shows the variation of SID accuracy of the combination with α for the same handset condition. It is clear that, for all $0 < \alpha < 1$, the combination gives a performance better than that of the classifier trained using only the MFCC. Also note that, for an optimal α of 0.67, the combination gives a maximum accuracy of 84.5% (an absolute accuracy improvement of about **12%** over that given by the MFCCs alone) which shows that the DCTILPR features capture speaker information missed out by the MFCCs. This is most probably because the MFCCs are vocal tract-based features while the DCTILPR is a VS-based feature, and thus the DCTILPR provides supplementary speaker information present in the VS which is not captured by the MFCCs.

Figure 4.7 shows how the SID accuracy varies over α for the different handset condition. There is no performance improvement with the combination, and in fact the performance drops for $0 < \alpha < 1$. This behaviour is mostly because both the features individually have significant handset variability, and the combination of DCTILPR adds to the handset variability already present in the MFCCs, thus degrading the performance.

4.5.4 Results with the DFT of ILPR

We obtained the pitch synchronous DFT of ILPR in the same way as we obtained the DCT, but we considered only the DFT magnitudes as the feature vector. The first 12 coefficients (excluding the 0^{th}) were considered, as they correspond to the first 12 harmonics of the fundamental (we take 24 coefficients in the case of DCT since the frequency resolution of the pitch synchronous DCT is $F0/2$). Results in Table 4.6 show that the DFTILPR performs poorer than the DCTILPR on TIMIT (8.3% reduction in accuracy). Also, on YOHO, if the train and test data are taken from different sessions, the accuracy given by DFTILPR is lower than that given by DCTILPR. This indicates that the DFT magnitude captures less speaker specific information than is captured by the DCT and the DFT phase has the rest of the information. Since the DCT captures both the magnitude and phase implicitly, it captures more speaker specific information and gives better accuracy. But, the DFT magnitudes perform poorer than the DCT on NIST 2003 in the different handset case also.

Table 4.6: Performance comparison of DFTILPR and DCTILPR

Database	SID accuracy (%)	
	DFTILPR	DCTILPR
TIMIT	86.4	**94.6**
YOHO: test session 1	100	100
YOHO: test session 2	72.5	**80.4**
YOHO: test session 3	67.4	**73.9**
YOHO: test session 4	59.4	**72.5**
NIST 2003: same handset	59.1	**71.8**
NIST 2003: different handset	14.5	**18.2**

When there is handset variability, since we expect a feature based only on the magnitude to perform better than a feature which captures the phase also, this is a surprising result. This may be because of the effect of the roll-off (especially the low frequency roll-off) of the 300 $Hz - 3.3$ kHz telephone quality bandpass filter on the DFT coefficients, but it has to be investigated in greater detail.

4.6 Speaker verification experiments on short test utterances

[1]One of the major bottlenecks in applying speaker recognition technology is the availability of data; at least 10-15 s of test data is required to get a reasonable performance. Results reported in [54] for short duration test data (<10 s) show that the performance drops significantly even though sufficient data is used for training. Since we can get enough training data in practice, we consider the case of limited (<10 s) test data, and test how the DCTILPR features and a combination of MFCC and DCTILPR features perform. The motivation for

[1]This work was in collaboration with Prof. S. R. M. Prasanna and Rohan Kumar Das from IIT Guwahati

doing so stems from the hypothesis that the VS-based features do not depend much on phonetic content, whereas the robustness of vocal tract-based features depends on the amount of phonetic content that they capture from a particular utterance. Since short test utterances contain very less phonetic content, VS-based features may result in a significant performance improvement when they are combined with vocal tract-based features.

4.6.1 The baseline i-vector system

The MFCC-based i-vector system for speaker verification (SV), which is the state-of-the-art SV system [7], is considered here as the baseline. A brief description of the system is given below.

Figure 4.8 shows the block diagram of the baseline system, which is reproduced from [7]. With a frame size of 20 ms and a shift of 10 ms, 13-dimensional MFCC features and their first and second order derivatives are extracted for each of the frames, thus making up a 39-dimensional feature vector. A gender-independent universal background model of 1024 mixture components is built using the development data provided in the NIST 2003 database, and it is adapted to the train and test data to obtain sufficient statistics, as proposed in [55]. The total variability matrix (T-matrix), which captures all the variability in the data, is learnt from the development data, and is used to project the train and test data to a lesser (400) dimensional space called the total variability space. These projections are the identity vectors (i-vectors), which are compact speaker representations. Thus, we have a train i-vector for each speaker and a test i-vector for each test utterance, and their variability (speaker and channel variability) is compensated by applying linear discriminant analysis (LDA) [56] and within-class covariance normalization (WCCCN) [57]. Scores are obtained by cosine distance-based scoring, and if the score crosses a particular threshold, the claim is accepted; otherwise, it is rejected.

Table 4.7 shows the SV performance of the baseline system, with the equal error rate (EER) and the decision cost function (DCF) as the performance metrics. We notice that the performance improves after compensation in all the cases, and the performance degrades as the test utterance duration reduces.

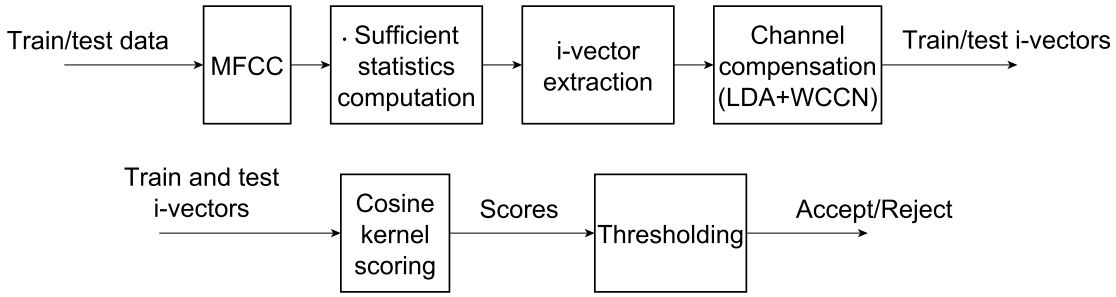


Figure 4.8: Block diagram of baseline i-vector system (taken from [7])

Table 4.7: Results of the baseline i-vector system on the entire 356-speaker NIST 2003 dataset using MFCC features for limited duration test segments.

Test Utterance Duration	**SV performance of the system**			
	Without compensation		**With compensation**	
	EER(%)	**DCF**	**EER(%)**	**DCF**
10 sec	8.85	0.1620	5.80	0.1090
5 sec	13.91	0.2631	10.52	0.1977
3 sec	19.82	0.3662	16.94	0.3100
2 sec	25.38	0.4784	22.31	0.4128

4.6.2 Performance of DCTILPR features

The baseline system is rebuilt with the DCTILPR features, and Table 4.8 gives its SV performance. The effect of compensation is seen more clearly than in Table 4.7, showing the larger variability of the DCTILPR features. Comparing Tables 4.7 and 4.8, the DCTILPR features do not perform as well as the MFCCs. This may be reasoned out to be mainly due to the larger handset variability shown by the DCTILPR (inferred from Table 4.5), a part of which may remain even after compensation.

4.6.3 Combination of MFCC and DCTILPR features

The baseline system is combined with the system with DCTILPR features at the score level, as described by equation 4.5, except that the likelihoods are replaced by the cosine kernel scores. Denoting the scores of the baseline (MFCC-trained) system, the DCTILPR-trained system and the combination by S_M, S_D and S_C respectively, we have

Table 4.8: Results of the i-vector system on the entire 356-speaker NIST 2003 database using DCTILPR features for limited duration test segments.

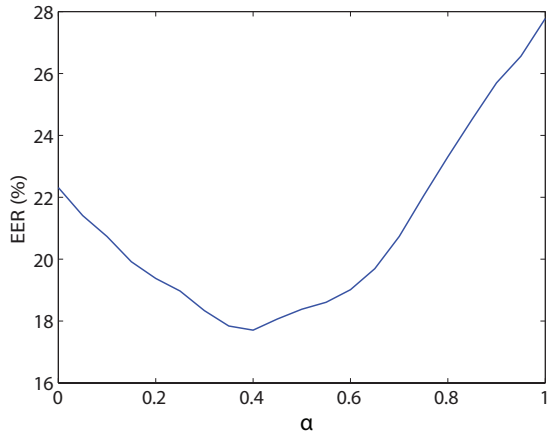
Test Utterance Duration	**SV performance of the system**			
	Without compensation		**With compensation**	
	EER(%)	**DCF**	**EER(%)**	**DCF**
10 sec	24.93	0.45	13.91	0.25
5 sec	27.59	0.52	18.65	0.35
3 sec	31.84	0.58	22.13	0.41
2 sec	34.73	0.65	27.78	0.52

$$S_C = \alpha S_D + (1 - \alpha) S_M \tag{4.6}$$

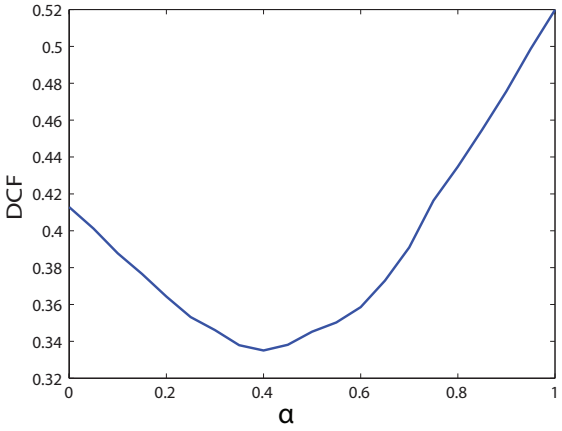
where α is again a scalar varying from 0 to 1. With the S_Cs obtained from Equation 4.6, the EERs and DCFs are computed for different test utterance durations, and are listed in Table 4.9. Table 4.9 shows that the performance improves significantly over that of the baseline for short utterances after fusion of the DCTILPR due to the additional speaker information present in it, which is not captured by the vocal tract features. α_{opt} varies between 0.15 and 0.4 for different cases, which shows that the DCTILPR features must be given a weightage in that range for optimal performance. It is observed that the improvement in EER over the baseline is more and more pronounced as the duration of the test data decreases (5.81%-5.33%= 0.48% in the 10 s case to 22.31%-17.71%= 4.6% in the 2 s case, with compensation). Also, α_{opt} increases as the duration of the test data decreases (0.15 in the 10 s case to 0.4 in the 2 s case, with compensation). Thus, the importance of the source feature increases as the test data duration decreases. Figures 4.9(a) and (b) show the plots of EER and DCF versus α for the 2 s case with compensation. We notice that the performance improves and reaches a maximum at $\alpha = 0.4$, thus showing that the VS-based features can significantly improve the SV performance of the MFCC-trained system for short test utterances.

Table 4.9: Performance of the proposed i-vector system for short test segments on the 356-speaker NIST 2003 database fusing DCTILPR and MFCC-trained classifiers at the score level. Improvement of the performance metrics over the baseline system are also listed.

Test	**Performance- proposed fusion system**						**Performance improvement over baseline**			
Utterance	**Without compensation**			**With compensation**			**Without compensation**		**With compensation**	
Duration	α_{opt}	**EER(%)**	**DCF**	α_{opt}	**EER(%)**	**DCF**	**EER(%)**	**DCF**	**EER(%)**	**DCF**
10 sec	0.25	7.90	0.1491	0.15	**5.33**	**0.0971**	0.95	0.0129	**0.48**	**0.0119**
5 sec	0.30	12.20	0.2290	0.30	**8.45**	**0.1567**	1.71	0.0341	**2.07**	**0.0380**
3 sec	0.30	16.98	0.3213	0.40	**12.46**	**0.2325**	2.84	0.0449	**4.48**	**0.0775**
2 sec	0.30	23.08	0.4313	0.40	**17.71**	**0.3351**	3.07	0.0471	**4.60**	**0.0777**



(a) EER vs α



(b) DCF vs α

Figure 4.9: EER and DCF versus α for the fusion of the DCTILPR and MFCC-trained classifiers on the 356-speaker NIST 2003 database.

Chapter 5

Conclusion and future work

A new non-parametric characterization of the VS by the DCT is proposed and explored. The ILPR is used as an estimate of the VS, and, since the DCT is known to be sensitive to start and end point shifts, the pitch synchronous DCT is used to characterize the VS. The pitch synchronous DCT coefficients are interpreted as levels of the various harmonics of the VS, and are used to test whether a speaker's voice can be characterized by the harmonics of the VS.

Visual observations of ILPR waveforms and the corresponding pitch synchronous DCT coefficients show that the DCT coefficients indeed capture speaker information, with the coefficients changing from speaker to speaker. An examination of the scatter plots of the DCT coefficients reveals that the features cluster in multiple clusters for a single speaker, showing that there is reasonable intra-speaker variability too. Nevertheless, an examination of scatter plots for different speakers shows that the distribution of the DCT coefficients may vary from speaker to speaker, and hence GMMs are used to capture this distribution.

By conducting SID experiments on TIMIT with the DCTILPR as the feature vector, it is found that diagonal covariance matrices outperform full covariance matrices, a result consistent with the findings of previous studies. It is also found that mixtures of 16 Gaussians best capture the feature distribution.

SID experiments on the TIMIT and YOHO databases show that most of the speaker information is captured in the first 12 DCT coefficients (6 harmonics) and 24 DCT coefficients (12 harmonics) give the maximum accuracy. A visual observation of the reconstructed ILPRs

by varying the number of DCT coefficients used for reconstruction shows that the gross shape of the ILPR is captured by the first 16 DCT coefficients, thus conclusively showing that most of the speaker information is contained in the gross shape of the VS. Surprisingly, it is found that the very fine structure of the VS (captured by harmonics higher than the 12^{th}) does not contain speaker information, but may instead lead to more confusion among speakers.

The SID performance of the DCTILPR is found to be comparable to the other proposed VS-based features. On the YOHO database, the DCTILPR is found to perform better than the state-of-the-art VS-based feature, namely the DSM glottal signatures. However, both the DCTILPR and the DSM features are found to suffer from session variability. On the NIST 2003 database, the DCTILPR is observed to provide significant supplementary speaker information when combined with the MFCCs. However, it is also observed to suffer from handset variability.

The DFT coefficients of the ILPR are also used as features for SID, and comparing its performance with that of the DCTILPR, it is seen that the DCTILPR performs consistently better than the DFTILPR, which indicates that there is significant speaker information in the phase of the VS. But, the DCTILPR performing better than the DFTILPR even when there is handset variability is counter-intuitive, and this issue needs to be examined in more detail.

Considering short test utterances on the NIST 2003 database, it is shown that a combination of MFCC and DCTILPR-trained systems performs significantly better than the MFCC-trained system, showing that the DCTILPR features can be deployed along with the MFCCs in a realistic scenario where the test data is $< 10s$ in duration.

The session and handset variabilities associated with the DCTILPR are its main shortcomings. However, they may be compensated to some extent using techniques like LDA and WCCN, as shown by SV experiments on the NIST 2003 database. Of course, if a new feature is found that does not suffer from these variabilities, we do not need any compensation techniques, but nature does not reveal itself so easily, and it is ideally towards this end that speaker recognition research must advance.

The ILPR may have phonetic variability due to improper formant cancellations, and thus a more sophisticated inverse filtering technique may be required. At the basic level, a formal

speech science study on the ILPR and its phonetic variability is needed, which may lead to an improved method of inverse filtering.

Though the proposed DCT-based characterization of the VS is motivated from the SID perspective, it can be used to mine any other information, e.g, prosodic information, which resides in the VS. Also, since it is a general characterization, it can be applied to studies/applications like speech synthesis, voice pathology, etc., where the VS is of use. Studies which test the usefulness of the DCTILPR in these applications are needed.

Finally, the 'different' speaker information provided by the DCTILPR, which is not present in the MFCCs, needs to be investigated in more depth. This may lead to more insight into the basic question - where/what is speaker information in speech?

Appendix A

Interpretation of the DCT coefficients as harmonics

Consider a single pitch period of length N of the ILPR shown in Figure A.1 (left panel), denoted by $v(n)$. Also consider its even symmetric extension shown in Figure A.1 (right panel), denoted by $v_{es}(n)$. We have a relation between the N-point DCT of $v(n)$ and the $2N$-point DFT of $v_{es}(n)$ [58]. Specifically,

$$N\text{−}point\ DCT\ of\ v(n) = first\ N\ points\ of\ the\ 2N\text{−}point\ DFT\ of\ v_{es}(n) \tag{A.1}$$

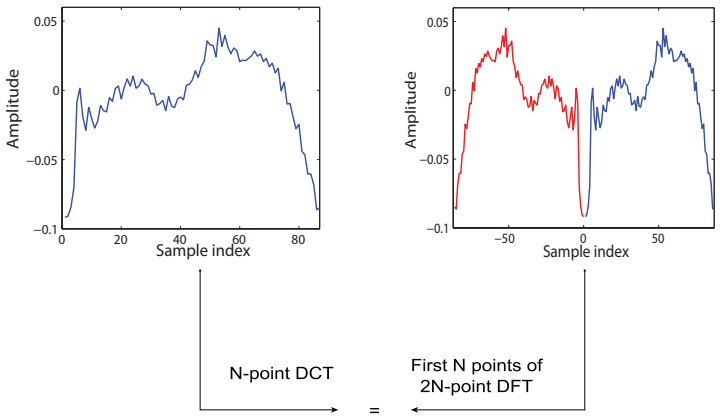


Figure A.1: Relation between DCT and DFT of the even symmetric extension

Now, the DFT coefficients of $v_{es}(n)$ are related to the discrete time fourier series (DTFS) coefficients of its periodicized version shown in Figure A.2 (right panel) [58], denoted by $v_{esp}(n)$. Specifically,

$$2N\text{-}point\ DFT\ of\ v_{es}(n) = DTFS\ of\ v_{esp}(n) \tag{A.2}$$

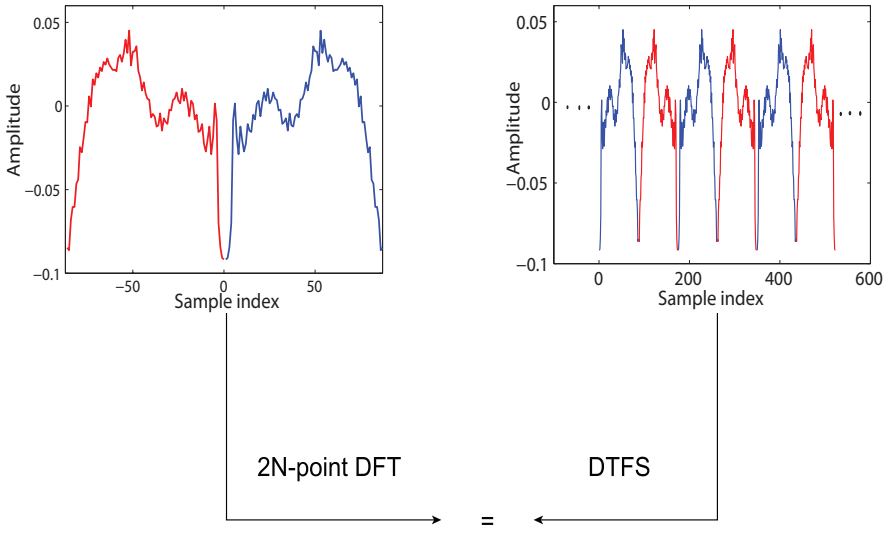


Figure A.2: Relation between DFT and DTFS of the periodicized version

From Equations A.1 and A.2, we can see that the DCT coefficients of $v(n)$ are the DTFS coefficients, or in other words, harmonics of $v_{esp}(n)$. Though $v_{esp}(n)$ is not the actual VS, it is derived from one period of the VS, and hence in this sense we can interpret the DCT coefficients as representing the harmonics of the VS.

Appendix B

Resynthesis of the speech signal by varying the number of retained DCT coefficients of the ILPR

In section 3.3.2, we saw that the reconstruction of the ILPR gets closer to the original, as the number (M) of DCT coefficients retained for reconstruction is increased. To observe the effect of changing M on the resynthesized speech signal, a similar experiment is conducted by resynthesizing the speech signal by varying M.

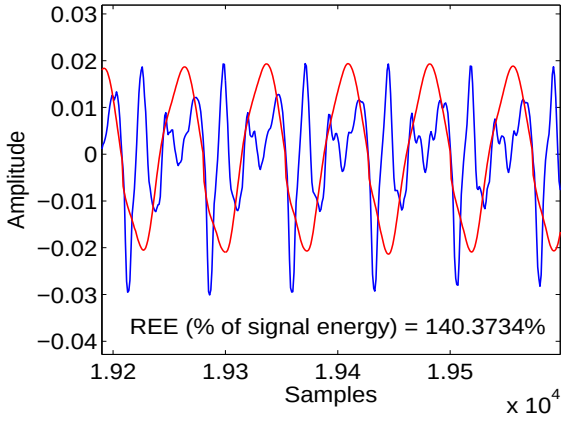
B.1 Resynthesis of speech from ILPR by varying M

The ILPR is extracted from the speech signal, as discussed in section 3.2.3. Since the LP coefficients are also obtained in the process of obtaining the ILPR, the speech signal can be resynthesized.

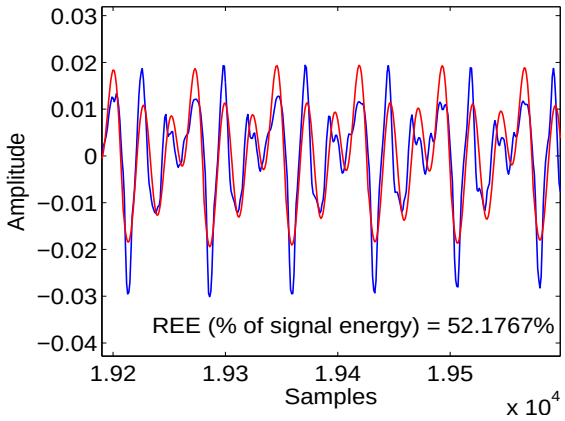
The DCT coefficients of the ILPR are obtained pitch synchronously, and the ILPR is reconstructed by varying M, as in section 3.3.2. Since, during analysis, the LP coefficients have been obtained from three consecutive pitch periods of the speech signal, the DCT is applied on three consecutive pitch periods of the ILPR. These three consecutive periods are reconstructed by varying M, and they are forward filtered using the LP coefficients obtained in the ILPR extraction process. Only the middle period of the resynthesized speech is

retained, and the process is repeated, shifting the analysis frame by one pitch period.

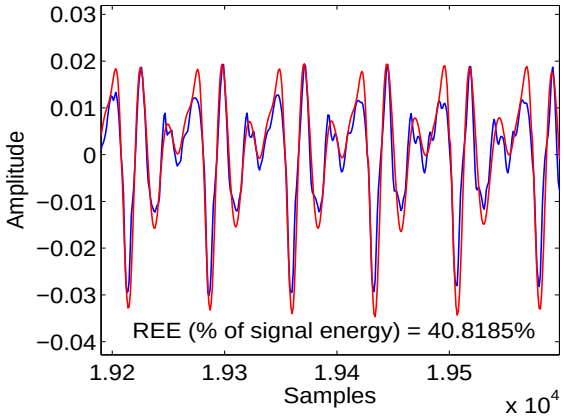
For a TIMIT speaker, Figure B.-1 shows the resynthesized speech signals by varying M in the range 10-60. The resynthesis error energy (REE) as a percentage of the original signal energy is also given. It can be easily seen that the speech signal is resynthesized better, as M increases. The resynthesized speech signal with $M = 30$ (Figure B.-1(c)) captures the gross time-domain shape of the speech signal. However, for $M > 40$, the resynthesized speech is almost the same as the original, showing that the higher coefficients capture the tiny detail present in the high frequency components. This is reflected in the REE, which decreases as M increases.



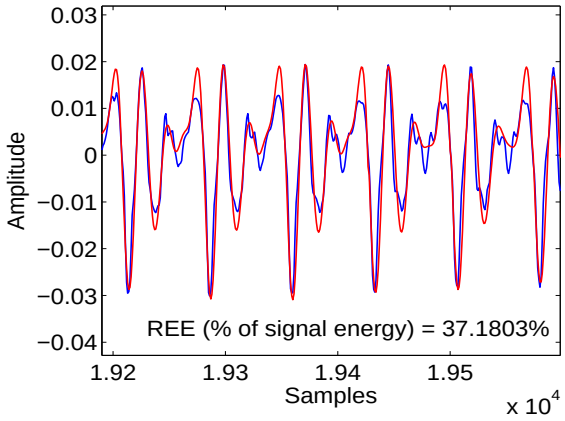
(a) Speech signal (blue) and its reconstruction (red) with the first 10 DCT coefficients of the ILPR



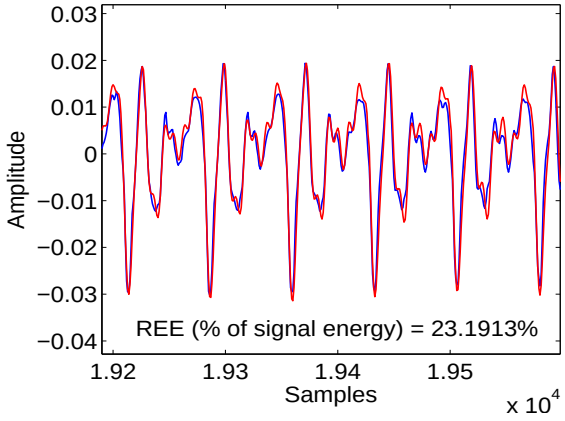
(b) Speech signal (blue) and its reconstruction (red) with the first 20 DCT coefficients of the ILPR



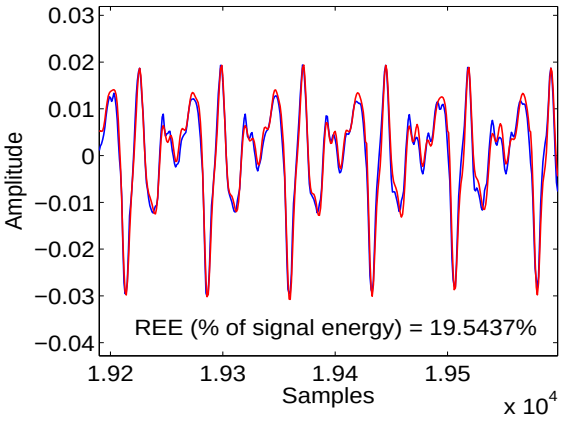
(c) Speech signal (blue) and its reconstruction (red) with the first 30 DCT coefficients of the ILPR



(d) Speech signal (blue) and its reconstruction (red) with the first 40 DCT coefficients of the ILPR



(e) Speech signal (blue) and its reconstruction (red) with the first 50 DCT coefficients of the ILPR



(f) Speech signal (blue) and its reconstruction (red) with the first 60 DCT coefficients of the ILPR

Figure B.-1: Voiced speech signal and its reconstruction from the truncated DCT of the ILPR, for varying number of DCT coefficients retained. The resynthesis error energy (as a percentage of the original signal energy) is also given.

Bibliography

[1] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2-3, pp. 109–118, 1992. (Cited on pages viii, 2 and 12.)

[2] ——, "Glottal inverse filtering analysis of human voice production-A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–651, 2011. (Cited on pages viii, 8, 9 and 10.)

[3] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 569–586, 1999. (Cited on pages viii, 6, 13, 14, 15, 16, 36 and 47.)

[4] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *Proc. ICASSP*, 2008, pp. 4821–4824. (Cited on pages viii, 6, 16, 46 and 47.)

[5] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, pp. 968–981, 2012. (Cited on pages viii, 6, 17, 18 and 50.)

[6] T.V. Ananthapadmanabha, "Acoustic factors determining perceived voice quality," in *Vocal fold physiology: voice quality control.* O.Fujimura and M. Hirano, Eds., San Diego, Cal.: Singular publishing group, ch. 7, 1995, pp. 113–126. (Cited on pages viii, 4, 6, 12, 21 and 22.)

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front–end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011. (Cited on pages ix, 57 and 58.)

[8] G. Fant, *Acoustic Theory of Speech Production*. Mouton De Gruyter, 1970. (Cited on pages 1, 20 and 21.)

[9] M. Rothenberg, "Acoustic interaction between the glottal source and the vocal tract," in *Vocal fold physiology*. K. N. Stevens and M. Hirano, Eds., U. of Tokyo Press, 1980, pp. 305–328. (Cited on page 4.)

[10] I. Titze and J. Sundberg, "Modeling source-filter interaction in belting and high-pitched operatic male singing," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1530–1540, 2009. (Cited on page 4.)

[11] D.Y. Wong, J.D. Markel, and A.H.Gray Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, pp. 350–355, 1979. (Cited on pages 4, 10 and 11.)

[12] T.V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Commun.*, pp. 167–184, 1982. (Cited on page 4.)

[13] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010. (Cited on page 4.)

[14] D.H. Klatt, "Review of text–to–speech conversion for English," *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 737–793, 1987. (Cited on page 6.)

[15] I. Karlsson, "Modelling voice variations in female speech synthesis," *Speech Commun.*, vol. 11, no. 4–5, pp. 491–495, 1992. (Cited on page 6.)

[16] J. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio Electroacoust.*, vol. 21, no. 2, pp. 298–305, 1973. (Cited on page 6.)

[17] P. Alku, H. Tiitinen, and R. Naatanen, "A method for generating natural–sounding speech stimuli for cognitive brain research," *Clinical Neurophysiology*, vol. 110, pp. 1329–1333, 1999. (Cited on page 6.)

[18] Y. Koike and J. Markel, "Application of inverse filtering for detecting laryngeal pathology," *Annals of otology, rhinology and laryngology*, vol. 84, pp. 117–124, 1975. (Cited on page 6.)

[19] P. Howell and M. Williams, "Acoustic analysis and perception of vowels in children's and teenagers' stuttered speech," *J. Acoust. Soc. Am.*, vol. 91, no. 2, pp. 1697–1706, 1992. (Cited on page 6.)

[20] D.G. Childers, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991. (Cited on pages 6, 11 and 12.)

[21] D.W. Farnsworth, "High speed motion pictures of the human vocal cords," *Bell Labs. Rec.*, vol. 18, pp. 203–208, 1940. (Cited on page 6.)

[22] S.R. Mahadeva Prasanna, C.S. Gupta, and B. Yegnanarayana, "Extraction of speaker–specific excitation information from linear prediction residual of speech," *Speech Commun.*, vol. 48, pp. 1243–1261, 2006. (Cited on pages 6 and 12.)

[23] H. Pulakka, P. Alku, S. Granqvist, S. Hertegard, H. Larsson, A.M Laukkanen, P.Ake Lindestad, and E. Vilkman, "Analysis of the voice source in different phonation types: Simultaneous high-speed imaging of the vocal fold vibration and glottal inverse filtering," in *Proc. 8th International Conference on Spoken Language Processing*, 2008. (Cited on page 8.)

[24] R. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Am.*, vol. 31, no. 6, pp. 667–677, 1959. (Cited on page 10.)

[25] M. Rothenberg, "A new inverse filtering technique for deriving the glottal airflow waveform during voicing," *J. Acoust. Soc. Am.*, vol. 53, no. 6, pp. 1632–1645, 1973. (Cited on page 10.)

[26] A. Krishnamurthy and D. Childers, "Two–channel speech analysis," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, pp. 730–743, 1986. (Cited on page 11.)

[27] G. Fant, J. Liljencrants, and Q. Lin, "A four–parameter model of glottal flow," *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 26, pp. 1–13, 1985. (Cited on page 12.)

[28] T.V. Ananthapadmanabha, "Acoustic analysis of voice source dynamics," *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 25, no. 2–3, pp. 1–24, 1984. (Cited on pages 12 and 24.)

[29] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter – a new method for quantification of the glottal flow," *Speech Commun.*, vol. 22, pp. 67–79, 1997. (Cited on page 12.)

[30] I. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *J. Acoust. Soc. Am.*, vol. 91, no. 5, pp. 2936–2946, 1992. (Cited on page 12.)

[31] D. Pati and S. R. M. Prasanna, "Speaker recognition from excitation source perspective," *IETE Tech. Review*, vol. 27, no. 2, pp. 138–157, 2010. (Cited on page 12.)

[32] B. Yegnanarayana, K. Sharat Reddy, and S.P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proc. ICASSP*, vol. 1, 2001, pp. 409–412. (Cited on page 12.)

[33] J. E. Dennis Jr, D. M. Gay, and R. E. Welsch, "Algorithm 573 nl2sol-an adaptive nonlinear least–squares algorithm," *ACM Trans. Math. Softw.*, vol. 7, pp. 369–383, 1981. (Cited on page 14.)

[34] D.A. Reynolds and R.C. Rose, "Robust text–independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995. (Cited on pages 15, 40 and 46.)

[35] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 34–43, 2007. (Cited on page 16.)

[36] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech*, 2009. (Cited on page 17.)

[37] T. Drugman, G. Wilfart, and T. Dutoit, "Eigenresiduals for improved parametric speech synthesis," in *Proc. EUSIPCO*, 2009. (Cited on page 17.)

[38] P.B. Pati and A.G. Ramakrishnan, "Word level multi–script identification," *Pattern Recognition Letters*, vol. 29, pp. 1218–1229, 2008. (Cited on page 25.)

[39] K.G. Aparna and A.G. Ramakrishnan, "A complete Tamil Optical Character Recognition system," in *Proc. Fifth IAPR Workshop on Document Analysis Systems DAS–02*, 2002, pp. 53–57. (Cited on page 25.)

[40] V. Britanak, P.C. Yip, and K.R. Rao, *Discrete cosine and sine transforms-General Properties, Fast Algorithms and Integer Approximations*. Academic Press, 2006. (Cited on pages 25, 26 and 46.)

[41] A.A. Livshin and X. Rodet, "Musical instrument identification in continuous recordings," in *Proc. of DAFX*, 2004, pp. 222–227. (Cited on page 26.)

[42] ——, "The significance of the non-harmonic 'noise' versus the harmonic series for musical instrument recognition," in *Proc. of the 7th International Society for Music Information Retrieval (ISMIR) conference*, 2006, pp. 95–100. (Cited on page 26.)

[43] J. Eggink and G.J. Brown, "Instrument recognition in accompanied sonatas and concertos," in *ICASSP*, 2004, pp. 217–220. (Cited on page 26.)

[44] R. Muralishankar, A.G.Ramakrishnan, and P. Prathibha, "Modification of pitch using DCT in the source domain," *Speech Commun.*, vol. 42, no. 2, pp. 143–154, 2004. (Cited on page 27.)

[45] A.P Prathosh, T.V. Ananthapadmanabha, and A.G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 12, pp. 2471–2480, 2013. (Cited on pages 28 and 29.)

[46] T.V. Ananthapadmanabha, A.P. Prathosh, and A.G. Ramakrishnan, "Detection of closure-burst transitions of stops and affricates in continuous speech using plosion index," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 460–471, 2014. (Cited on pages 28 and 30.)

[47] D.G. Childers and C. Ahn, "Modeling the glottal volume velocity waveform for three voice types," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 505–519, 1995. (Cited on page 32.)

[48] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006. (Cited on pages 40 and 46.)

[49] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977. (Cited on page 40.)

[50] J. A. Saghri, A. G. Tescher, and J. T. Reagan, "Practical transform coding of multispectral imagery," *IEEE SP Magazine*, vol. 12, pp. 32–43, 1995. (Cited on page 46.)

[51] W. Fisher, G. Doddington, and K. Goudie–Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99. (Cited on page 47.)

[52] J. Campbell, "Testing with the YOHO CD–ROM voice verification corpus," in *Proc. ICASSP*, 1995, pp. 341–344. (Cited on page 47.)

[53] NIST Multimodal Information Group, "2003 NIST Speaker Recognition Evaluation, Linguistic Data Consortium, Philadelphia." (Cited on page 48.)

[54] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i–vector based speaker recognition on short utterances," in *Proc. Interspeech*, 2011, pp. 2341–2344. (Cited on page 56.)

[55] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, *Speaker verification using adapted Gaussian mixture models*. Digital Signal Processing, 2000. (Cited on page 57.)

[56] R. Vogt, S. Kajarekar, and S. Sridharan, "Discriminant NAP for SVM speaker recognition," in *Proc. IEEE Odyssey: Speaker Lang. Recogn. Workshop*, 2008. (Cited on page 57.)

[57] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. International Conference on Spoken Language Processing*, 2006. (Cited on page 57.)

[58] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete–Time Signal Processing (2nd Edition)*. Prentice-Hall Signal Processing Series, 1999. (Cited on pages 64 and 65.)

Publications based on this thesis

- A. G. Ramakrishnan, B. Abhiram, and S. R. Mahadeva Prasanna, "A characterization of the voice source using pitch synchronous discrete cosine transform for speaker information", submitted to *JASA Express Letters* on May 2, 2014, and being revised based on review comments.

- Rohan Kumar Das, B. Abhiram, S. R. Mahadeva Prasanna, and A. G. Ramakrishnan, "Combining source and system information for limited data speaker verification", accepted for oral presentation in *Interspeech 2014*.