Analysis of Vowel Articulation and Perception Using Acoustic Parameters

Thesis Submitted in partial fulfilment
of the requirements for the degree of

Master of Science
in
Electronics and Communication Engineering
by Research

by

Haala Deeba Abbas
20162302
haaladeeba.abbas@research.iiit.ac.in

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

(Deemed to be University)
Hyderabad - 500 032, INDIA
MAY 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Analysis of Vowel Articulation and Perception Using Acoustic Parameters**" by Haala Deeba Abbas, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Suryakanth V Gangashetty

Dedicated to my parents
Dr. Mohammed Abbas Ali and Dr. Adib Maleka

Acknowledgements

To begin with, I would like to acknowledge and express my gratitude towards my supervisor Dr. Suryakanth V Gangashetty for his constant guidance through out the research. I also hold huge regards for his generosity, simplicity, patience and timely help he has been rendering through out my masters journey.

I would like to thank my parents for their constant belief in me, without whom it wouldn't have been possible to complete my thesis today. I have no enough words to describe their unconditional love and support.

I would like to extend my thanks to all my seniors and fellow labmates. Without their support, it wouldn't have been possible to finish this work with much ease. My special thanks goes to RaviShankar Prasad, Bhanu Teja Nellore, Sudarsana, Sushmitha for their contribution towards guiding me in every aspect needed.

I thank all my wonderful friends for always being there for me and for keeping me motivated all the time. Their persistent belief in me had always helped me to carry out my research work with much zeal, besides giving me the strength to tackle the obstacles encountered. A special thanks to Sushmitha and Ayushi for always being by my side.

A debt of gratitude is owed to Sai Krishna, for his support and creating a friendly environment at the time of me joining IIIT. His moral support had also kept me stress free to explore research. I would like to thank the faculty of IIIT for providing courses that are in accordance with present state-of-art.

I am greatly thankful to IIIT's Library staff and Administrative staff, who have helped in any direct or indirect way.

Last but not the least, I would like to thank God for countless blessings.

Abstract

Speech is the most basic and reliable form of communication. In signal processing, speech signal carrying the message is represented in the form of an acoustic waveform. Having an insight into acoustics of speech contributes to production and perception robustness. One of the best means to study this is by exploiting the acoustic correlates of speech sounds. In this thesis, work mainly centers on introduction of acoustic correlates of vowels and their application in speech production, perception and analysis. Study of acoustic correlates of speech sounds brings out a rigorous introduction to acoustics of speech. Vowels are sonorants and the most audible sounds of speech. All the three works in this thesis mainly focus on studies related to vowel sounds and use formants as the basic spectral cue. Minor work has been extended to nasals in the first study. Formant bandwidth, one of the crucial acoustic correlates, was extracted for oral and nasal tracts using vowel and nasal articulations respectively. The exploitation of acoustic correlates of vowels against speech performance among different categories of gender was conducted in the second study. Here, the computed formants are utilized to construct vowel space area used for the comparison. Also, the role of source-filter was analyzed in perception of vowels. This study too incorporated formants in the vocal tract response of the LP method. Studying the acoustic correlates correlates of speech sounds would greatly enhance the knowledge on acoustical aspects of speech useful in Acoustic Phonetics, Auditory Phonetics, Speech Analysis, Speech Synthesis, Speech Signal Processing, to understand and evaluate speech signal.

Contents

List of Figures

List of Tables

List of Abbreviations

dB decibels

VS Vowel Space

F1 First Formant

F2 Second Formant

F3 Third Formant

HE Hilbert Envelope

F0 Fundamental Frequency

VOP Vowel Onset Point

VEP Vowel End Point

VSA Vowel Space Area

GVS Great Vowel Shift

ATR Advanced Tongue Root

LP Linear Prediction

AM Amplitude Modulation

FM Frequency Modulation

ESA Energy Separation Algorithm

IPA International Phonetic Alphabet

EWAR Exponentially Weighted Autoregressive

SDLMS Second Derivative of the Log–Magnitude Spectrum

CLSM Clustered Line–Spectrum Modeling

GDF Group Delay Function

NGD Numerator Group Delay

DNGD Double-differenced NGD

HNGD Hilbert Envelope of the Numerator Group Delay

ZFF Zero Frequency Filter

ZTW Zero Time Windowing

IIR Infinite Impulse Response

DEGG Differenced Electro-glotto-graph

GCI Glottal Closure Instant

LPCs Linear Prediction Coefficients

DTFT Discrete-Time Fourier Transform

DFT Discrete Fourier Transform

ALS Amyotrophic Lateral Sclerosis

PTSD Post-Traumatic Stress Disorder

FCR Formant Centralization Ratio

LnVSA Logarithm Vowel Space Area

IPD Idiopathic Parkinson's disease

FCOG formant centre of gravity

IIIT-H International Institute of Information Technology, Hyderabad

USB Universal Serial Bus

Chapter 1

Introduction

One of the challenges in speech research is the determination of certain attributes that help in the perception of speech sounds. These attributes are called acoustic correlates or acoustic cues. Study of vowel sounds require knowledge of acoustics in speech. Hence, it becomes necessary to study the acoustic correlates of vowels. Vowel sounds are the result of vibrating vocal folds exciting the fixed vocal tract with quasi periodic pulses of air [7]. Vowels are sonorants and are produced with a relatively open vocal tract.

In past, [8] used F0, F1, F2, F3, intensity, and duration, [9] used amplitude and [10] used spectral shape features and formants as acoustic correlates of vowels. Studies from past found formant frequency as one of the effective features to represent vowels [11]. Also, vowel space has been used as an effective acoustic metric in assessing vowel quality, speech intelligibility.

Acoustic data include vowel duration, steady-state formant frequencies, spectrogram and two measures of dynamic formant movement. Let us have a look at these acoustic correlates of vowels. This thesis work majorly covers extensive work related to vowels. Minor work has been carried out on nasals. Therefore, this chapter covers information on acoustic correlates of vowels in detail and very less discussion is illustrated on production of nasals. The reason behind throwing light on nasal sounds is the work carried out on both vowels and nasals in Chapter 2.

This chapter is organized as follows: Section 1.1 gives a detailed explanation on acoustic correlates of vowels. Features of vowel sounds and categorization is discussed in Sec. 1.2. Section 1.3 lists the features of vowel quality. Section 1.4 explains the production of nasal sounds. Applications of this study are mentioned in Sec. 1.5. Section 1.6 explains the motivation behind this study. The objectives and scope of thesis is described in Sec. 1.7. Section 1.8 lists the organization of the thesis.

1.1 Acoustic Correlates of Vowels

- **Vowel Duration:** The interval between the Vowel Onset point (VOP) and the Vowel End Point (VEP) defines the duration of the vowel. Long vowels and short vowels have also been distinguished based on this acoustic cue. Also, the duration of the vowels vary as per their surroundings viz., isolated vowels or vowels bounded by consonantal context like CVC. Moreover, vowels bounded by consonants appear to be shorter in length, and they lengthen at syntactic boundaries; Also, stressed vowels shorten in the vicinity of unstressed syllables [12].

- **Fundamental Frequency:** During the production of a voiced sound, a fundamental frequency (F0) along with its harmonic components (due to vibration of vocal folds) is created. Higher vowels have higher fundamental frequency [13]. In other words, fundamental frequency also known as pitch, having auditory properties, being the main factor in speech perception, enables the listener to perceive speech sounds [1].

- **Formants:** The vocal fold excitation modified by the vocal tract results in formants (poles) and anti-formants(zeros) [14]. Formants are characterised by regions of high energy in the frequency spectrum of voiced regions of speech. Formants play an important role in classifying vowels. The first two formants (F1 and F2) are the most reliable and relevant acoustic parameters [15]. The movement of different articulators together with vocal tract configurations lead to generation of formants [16]. The way the vowels are categorized as per the formant frequencies is discussed in Sec. 1.2. Fig. 1.1 shows F1, F2 and F3 peaks for front (/i/, /a/) and back (/ɑ/, /u/) vowels. For front vowels, F1 and F2 are far apart whereas vice-versa for back vowels. The most important cues considered in speech synthesis and also in speech processing are F0 and formants.

- **Formant Amplitude and Formant Bandwidth:** Each formant frequency has an amplitude and bandwidth associated with it. The formant whose peak amplitude is A, then the formant bandwidth is given as the difference in frequency between two points on either side of the peak having an amplitude A/(sq. root of 2) (corresponds to 3 dB down from peak).

- **Spectrogram:** A spectrogram is a 3D representation displaying spectrum of a speech signal. It displays time, frequency and intensity of the dark bands. The dark bands seen from Fig. 1.2 shows first three formants frequencies (indicated by arrows) of the vowels. The regularly spaced vertical lines on the spectrogram indicate the vibration of the vocal folds. These lines are visible throughout a large part of the spectrogram across the vowel region. Each line visible is a result of the acoustic energy arising from the single movement of vibrating vocal folds. This helps in determination of the pitch. Also, the observations

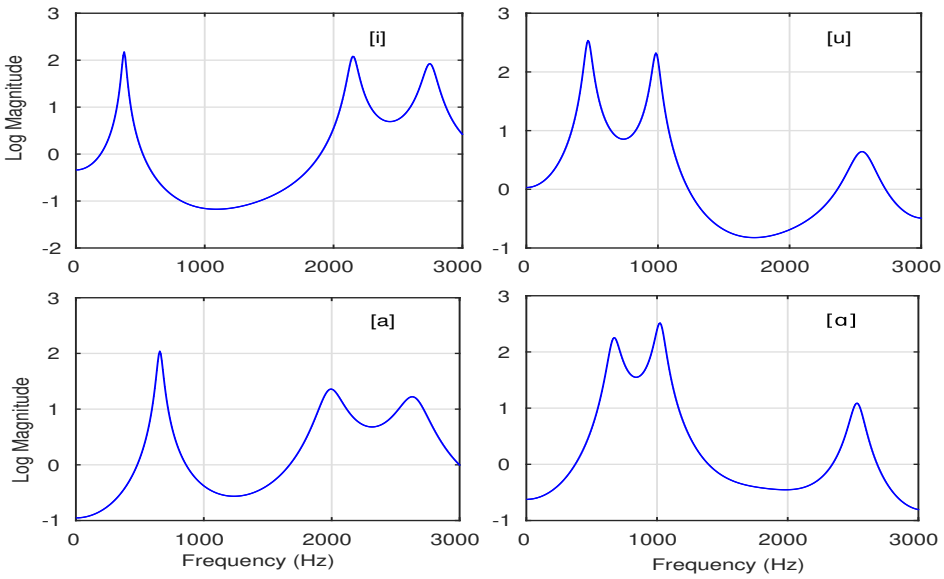

Figure 1.1: Spectra of front (/i/, /a/) and back (/ɑ/, /u/) vowels depicting F1, F2 and F3 frequency peaks

from the vertical striations on spectrogram aid in measuring pitch roughly. The pitch is higher when the vocal folds are close together, than when they are far apart. Fig. 1.2 displays spectrograms of words heed, hid, head, had, hod, hawed, hood, who'd spoken in British English by a male speaker. The horizontal axis displays intervals of 100ms. The vertical axis shows frequencies upto 4000 Hz. The first three formants shown by the dark bands horizontally are indicated by the arrows.

- **Vowel Space Area:** The formant frequencies of a vowel help in defining acoustic space [17]. This produces vowel space area (VSA), that is a 2-D area constructed using lines joining the coordinates of first two formant frequencies (F1) and (F2) [18], [19]. VSA helps in building the relationship between a speaker's contribution of his maximum time in articulating vowels and speech intelligibilty at the acoustic-perceptual levels. Thus VSA serves to be an important acoustic cue in studies related to vowels. Fig. 1.3 shows F1 frequencies plotted against F2 frequencies. The 3D representation (as can be seen from Fig. 1.4) of this gives the VSA. VSA is constructed using the Euclidean distance between F1 and F2 coordinates of the corner vowels /i/, /u/, /a/ (triangular VSA) or the corner vowels /a/ , /i/, /u/ and /ae/ (quadrilateral VSA) in F1-F2 plane [20]. These vowels are frequently selected as they are the most commom in human languages [21]. Also these vowels represent the extreme positions (which means maximum space allowed) in a talker's vowel working space and corresponding extreme formants are generated [22].

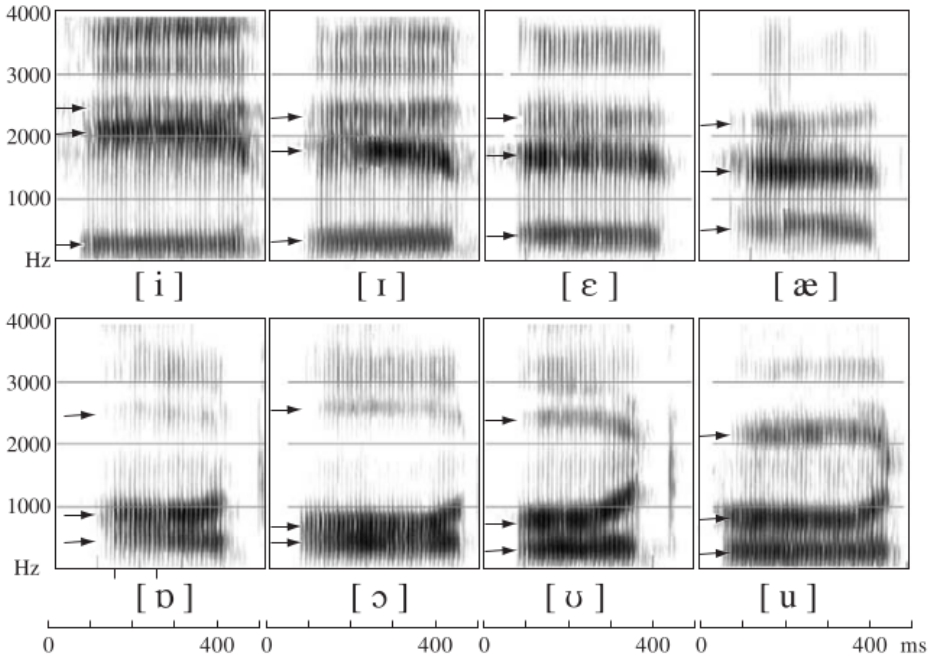

Figure 1.2: The spectrogram of the words heed hid, head had, hod, hawed, hood and who'd spoken by a male speaker of British English [1]

Acoustic correlates such as formants and VSA play a crucial role in the categorization of vowels [1]. Other acoustic cues such as vowel duration, steady state formants and dynamic formant movement [23] were used to assess vowel intelligibilty. Raising or lowering of these formant frequencies may enhance or reduce intelligibilty of vowels. Section 1.2 discusses about features of vowel sounds by discussing different categories of vowel sounds depending on how their production characteristics and also how their acoustic correlates vary.

1.2 Features of Vowels

Features of different vowel sounds vary as per their categorization. Acoustic correlate dependence is associated with categorization of vowels. For instance, formants F1 and F2 related to vertical (tongue height) and the horizontal (frontness/backness of tongue) positions of the tongue help in categorizing vowels. This relation is explained in next sub-sections respectively. Also, the reference vowels used in English and their features such as rhotacization, tenseness/laxness, tongue root advancement, nasality, vowel quality all are discussed in next sub-sections of this chapter.

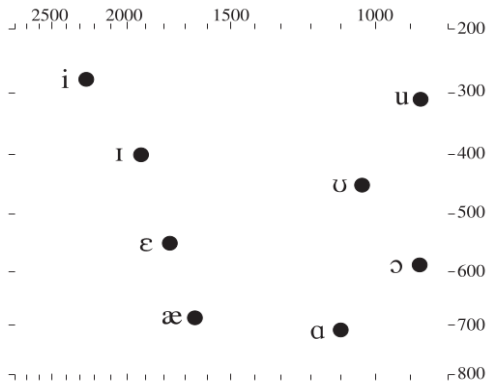

Figure 1.3: Formant chart showing the frequency of the first formant vertical axis) plotted against second formant (horizontal axis) for eight American English vowels [1]

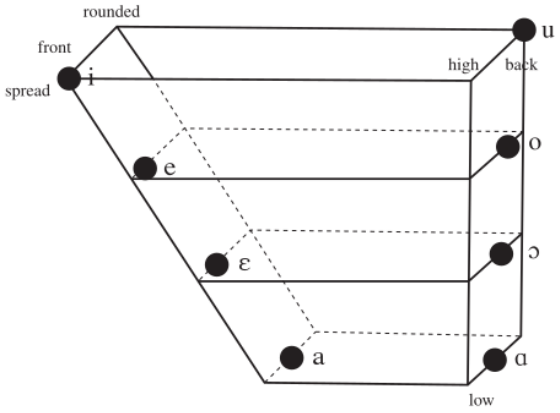

Figure 1.4: A three-dimensional representation of the vowel space, showing that the cardinal vowels fall on a plane that cuts across the space [1]

1.2.1 Cardinal vowels

To begin with vowel categorzation, we need to first describe the 18 reference vowels invented by Daniel Jones [24]. Vowels of a language can be best represented by a vowel chart [2]. Thus, these vowels are represented by a cardinal vowel chart as shown in Fig. 1.5. These vowels referred as 'cardinal vowels' indicate the extreme corners of the vowel chart within which the vocal tract moves. The reason behind using this vowel chart is that it becomes possible to describe vowels and compare vowel systems of one or more languages using these distinctive features. At the same time, these vowels do not correspond to any particular language, but serve as the nearest cardinal vowels possible in relation to the natural vowels produced [24]. As can be seen from the Fig. 1.5, all the vowels except 17 and 18 occupy the extreme edges of the vowel chart. This shows the maximum articulatory space utilised by a speaker.

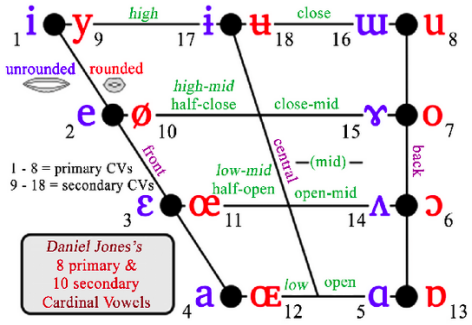

Figure 1.5: Cardinal Vowel chart [1], [2]

1.2.1.1 Primary Cardinal Vowels

Vowels numbered from 1 to 8 counter-clockwise (in Fig. 1.6), from the upper left corner.are the primary cardinal vowels. These are described as close front, mid-close front, mid-open front, open front, open back, mid-open back, mid-close back and close back respectively [24].

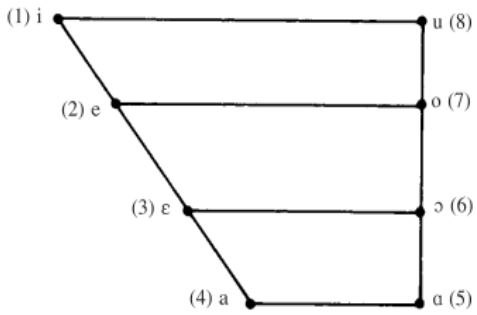

Figure 1.6: Primary Cardinal Vowels [1]

1.2.1.2 Secondary Cardinal Vowels

Vowels 9 to 16 shown in Fig. 1.7 are the secondary cardinal vowels. Their positioning is same as the primary cardinal vowels in the vowel chart, but the difference lies in the sound being less familiar to us, as the shape of lips is reversed. Lip rounding also plays a pivotal role in this vowel chart. For instance, vowels 9 to 13 are produced with rounded lips whereas vowels 14 to 16 with unrounded ones [24]. The remaining two close central vowels 17 and 18 are produced with unrounded and rounded lips respectively [24].

1.2.2 High & Low vowels/Close & Open Vowels

Formant F1 is the key factor in distinguishing high from low vowels [1]. High (close) vowels are produced when the tongue reaches the highest point (roof of the mouth) in the vocal tract,

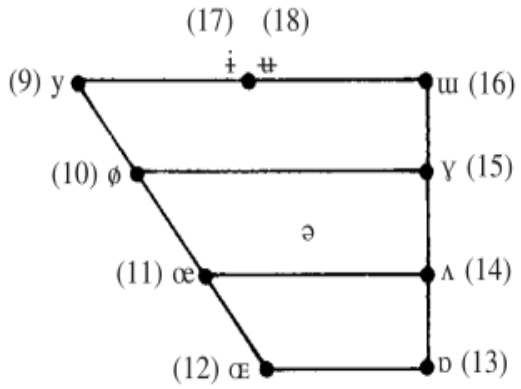

Figure 1.7: Secondary Cardinal Vowels [1]

and vowels in which the tongue height is at middle and low positions are known as mid and low (open) vowels respectively. This vertical position (closeness/openness) of the vowel is given by the first formant frequency F1. A higher vowel would be having a low F1, and vice-versa. As can be seen from the Fig. 1.8, F1 decreases as we keep going down from close high vowel 'i' to open low vowel 'a'.

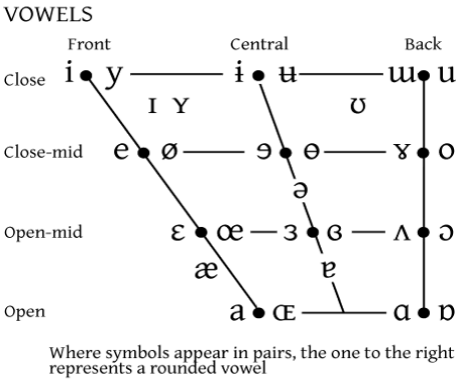

Figure 1.8: IPA Vowel Chart [1]

1.2.3 Front/Back vowels

F2 value depends on tongue advancement like how front or back the tongue is placed in the vocal tract [1]. Front vowels, that are, vowels produced with tongue placed at front of the vocal tract have higher F2 compared to back vowels, i.e., vowels produced with tongue placed at the back of the vocal tract. As can be seen from the chart shown in Fig. 1.8, vowels 'i', 'e', 'ɛ', 'æ' and 'a' are front vowels and vowels 'u', 'o', 'ʌ' and 'ɑ' are back vowels. F2 decreases as we keep moving from front vowel 'a' to back vowel 'ɑ'.'

Rounded/Unrounded vowels

Another vital articulatory gesture for distinction of vowels is the shape of lips. Usually, the position of the lips (rounded or spreaded) affects the third formant F3 [2]. Few vowels involve rounded lips, whereas few involve spread lips (in varying degrees) and few in neutral shape [25]. Vowels like 'o' in *road* and 'u' as in *true* are examples of rounded vowels. One example of vowel produced with extreme spreading of lips is 'i' in *sheep*.

1.2.4 Tense/Lax vowels

Tenseness/Laxness is another distinction for categorization of vowels. Vowels can be either tense or lax depending on the amount of muscle tension used to produce them. It also depends on the tendency of the vowel to glide from one articulatory position to other (as evident from the vowel chart). It also depends whether it lies in closed or open syllables, and its relative place of articulation [25]. Examples of tense vowels are /i,eɪ,a,ɔ,aʊ,u/ produced with mass muscle tension than the lax vowels /ɪ,ɜ,ʌ,æ, u/ [25]. The tense vowels are positioned less centrally due to the stretch caused by extra muscle tension, making them to occupy extreme peripheral positions in the vocal tract [25]. Also two important features of tense vowels are: They occur both in stressed open syllables, as in *day*, *shoe*, *row*, *zoo*, where the syllables do not end with a final consonant sound, and in stressed open syllables where the syllables terminate with a final consonant sound as such in *mean*, *cot*, *role*, *room*, *look* [25]. The other distinctive feature of tense vowels is their ability to glide. This may result in diphthongization of vowel sounds (discussed next). On other hand, lax vowels are produced with relaxed muscles, and hence they do not usually diphthongize as the tense ones do. Unlike tense vowels, they occur only in closed syllables like in *rim*, *let*, *run*, *put*. In other words, a lax vowel always needs a consonant to close the stressed syllable [25].

1.2.5 Diphthongs

The combination of two vowel sounds within a single syllable makes a diphthong. One remarkable event from the past known as 'Great Vowel Shift' (GVS) (can be seen from Fig. 1.9) also results in Diphthongization. This phenomenon occurs as the tongue glides from one vowel position to other resulting in two vowels getting combined, such as the transition from 'a' to 'i' resulting in 'ai' (as in bide). Diphthongs too are used for distinguishing vowel sounds by most of the British English speakers [1]. Diphthongs used in British English are as follows: /ɪə/ in words like *beer*, *hear*; /ʊə/ in words like *tour*, *pour*; /eə/ in words like *pair*, *hair*; /eɪ/ in words like *say*, *pay*, *ray*; /ɔɪ/ as in *toy*, *boy*; /aɪ/ as in *sky*, *shy*, *buy*; /əʊ/ as in *show*, *boat*, *coat*; and /aʊ/ as in *foul*, *houl*, *mouth*. Thus, it is evident that diphthongization of vowels lead to formation of long vowels.

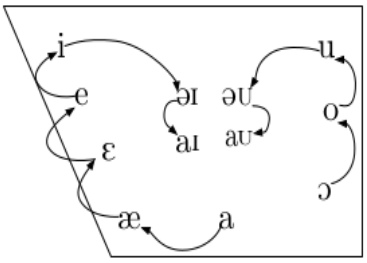

Figure 1.9: Vowel Space illustration of Great Vowel Shift [3]

1.2.6 Rhotacized vowels

The vowels produced with rhotacization are called rhotic vowels. These vowels are said to be r-colored and often referred to as retroflex vowels as their production involves retroflexion similar to that of consonants. As we know, retroflexion involves the tip of the tongue raised towards the back of hard palate in a repetitive fashion in consonants. Similarly, rhotic vowels can be produced in two distinct ways: Either by curling the tip of the tongue against hard palate, or by keeping the tip down such that a high bunched position of the tongue is achieved. Both the positions result in sound having a similar auditory effect. In vowels from many forms of American English, the rhotic vowels evolve when the vowels are followed by [r]. Examples are *board*, *bared*, *hour*. Therefore, the individual quality of the vowel is not compromised as the r-coloration of the vowel is not evident at the beginning of the vowel [1]. But in few cases, the whole vowel is rhotacized as such in words like *fur*, *first*, *nerd*, *learn*. In all these, the mid-central vowels have added rhotacization. Third Formant (F3) is associated with rhotacization. Rhotic vowels have a lower F3 [1]. Overall, rhotacization of the vowel refers to how it sounds [1].

1.2.7 Advanced Tongue Root vowels

As the name itself implies, the vowels that involves the roof of the tongue being drawn forward with lowering of larynx, causing the part of vocal tract to be drawn forward and making it enlarged are called Advanced Tongue root [ATR] vowels, commonly denoted by [+ATR] vowels, and if there is no advancement involved, then it is simply denoted by [-ATR] vowels. The advantage of studying the contribution of such vowels involves the fact that the acoustic correlates such as formants vary according to the tongue involvement, which in turn also affects the VSA. Vowels i, u, e and o are [+ATR] vowels whereas vowels ɪ, ʊ, ɛ, ɔ and a are [-ATR] vowels [1].

1.2.8 Nasal Vowels

Vowels whose production is accompanied with lowering of velum, so as to allow air escape through the nose are called nasalized vowels or simply nasal vowels [1]. Here the air passes through both the oral and nasal tracts. One such example is [æ] as in *man* [*maen*], because the vowel is terminated by a nasal consonant 'n'.

1.3 Features of Vowel Quality

Altogether, if all these vowel categorizations are summarized, then they can be illustrated with the help of Table. 1.1 (taken from [1]).

Quality	**Correlates**
height	F1
backness	difference between F1 and F2
rhotacization	F3
rounding	lip position
ATR	width of pharynx
nasalization	position of velum

Table 1.1: The features of Vowel Quality [1]

1.4 Production of Nasals

Nasals in English 'm', 'n' and 'ŋ' belong to the family of sonorant consonants that involves vibration of vocal folds during their production [26]. One remarkable aspect in the production of nasals is the primary contribution of a structure called velum (also known as soft palate). Velum (as can be seen from the Fig. 1.10) serves as the switch between nasal and non-nasal sounds, and gets shut off at the entrance of nasal cavity to produce sounds other than nasals [27]. Nasals are produced by lowering the velum while air is allowed to flow through the nasal cavity besides constriction being made in the oral cavity. Also, the velopharyngeal port is opened with lowering of velum, resulting in coupling of both oral and nasal tracts. This gives rise to a larger production cavity.

Nasals are distinguished by the point of constriction made at different articulatory positions. Like, the point of constriction for 'm' is at lips, resulting it to be a bilabial nasal. For 'n', the point of constriction occurs at alveolar ridge, giving rise to an alveolar nasal and for 'ŋ', it occurs at the forward of velum. The vocal tract configurations for the nasals 'm', 'n' and 'ŋ' can be

seen from Fig. 1.11. The spectral characteristics of nasals are determined by these points of constriction [28].

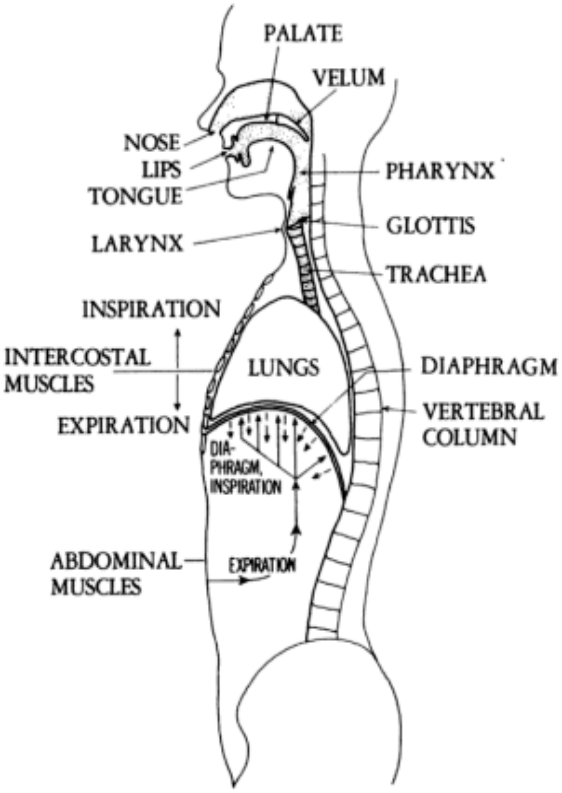

Figure 1.10: A schematic representing speech production mechanism [4]

The coupling of both nasal and oral tracts introduces poles and zeros in the low frequency region of coupled nasal tract [29] leading to resonances.

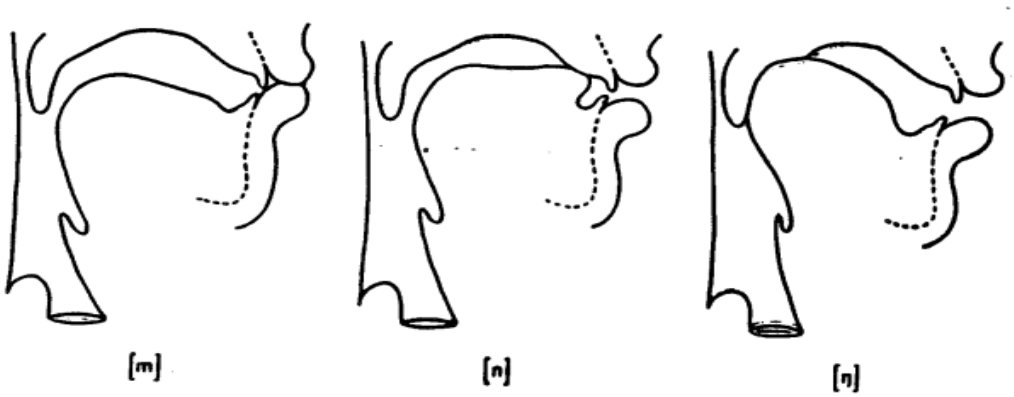

Figure 1.11: The vocal tract configurations for nasal consonants 'm','n' and 'ŋ' [5].

Nasals in speech are mostly identified from their behavior of exhibiting a low frequency pole and a following zero. Also, spectral cues such as their resonance locations and respective bandwidths also help in their identification [30], [31], [32], [33] and [34]. Although the oral cavity is constricted at front, it still remains acoustically coupled with pharynx behind, giving rise to antiresonances or zeros. The nasal cavity has a relatively large surface area to cross-sectional area ratio, resulting in resonances that are broader spectrally or highly damped. The larger this ratio, the more the heat conduction and viscous losses are. The nasal cavity is longer than the oral tract, and also exhibits a higher impedance. This behavior results in an increase in the bandwidth of the resonances associated with the nasal tract in the spectrum. Also, the dissipative energy losses within the nasal cavity are proportional to the shape of the air cavities inside the nasal tract. This results in high formant bandwidths of nasal sounds [35].

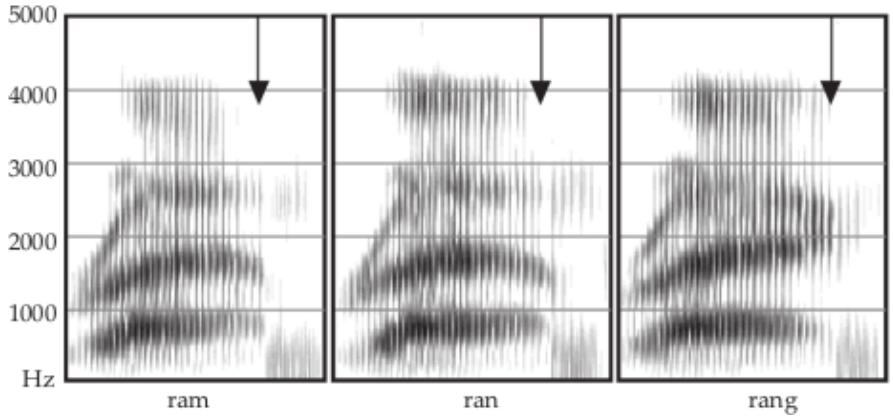

Figure 1.12: Spectrograms of nasals at the ends of the words ram, ran, rang. The arrows mark the onsets of the nasals [2]

As can be seen from the spectrograms of the three nasals in words ram, ran and rang in Fig. 1.12, the F1 bands of all the three nasals appear fainter, indicating less energy compared to the preceding vowel. Again, F2 of three nasals exhibit same pattern, which is more lighter comparatively. The main reason behind this lighter spectral band prominence of nasals is due to the effect of nasal and oral tracts where, combination of resonances and anti-resonances take place. Thus, it would be interesting to study the bandwidth of such resonances involved in nasal tract which would be discussed in next chapter.

1.5 Applications of the study

Study of vowel sounds and their acoustic correlates finds applications in Speech Analysis, Speech Synthesis and Speech disorders. Resonance bandwidth obtained across nasal and oral tracts finds applications in Acoustics of Speech, modeling the vocal tract length and various

other parameters accordingly. Studying vowel space provides a detailed insight into clinical speech research, vowel production, vowel reduction and disorders, vowel recognition, speech pathology.

1.6 Motivation of the study

Vowel sounds play a significant role in speech processing. This is because vowels are the most sonorous sounds, form the nucleus of a syllable. It becomes very easy to speak, as vowels are produced with an open mouth, thereby making communication effortless. Also, vowels are the best means to express emotions. Without vowels, it would become very difficult for a speaker to convey emotions to the listener. The advantage in studying vowel sounds is attributed from the fact that these sounds are sonorants and they always possess the greatest energy in the speech signal, which further helps in exploration of their acoustic correlates. Also, this helps in knowing the characteristics of vowels. The application of these correlates to the thesis work highlights the contribution of vowels in assessing ariculatory space in speech production and their role in the perception of speech modeled through source-system separation system.

1.7 Objectives and Scope of the thesis

The main objectives of this thesis work are:

- To study acoustic correlates of vowels.
- To extract formant bandwidth for oral and nasal tracts.
- To compare the speech performance among different genders by exploiting acoustic correlates of vowels such as formants and vowel space area.
- To understand the role of excitation source and vocal tract system in the perception of vowels.

1.8 Organization of the thesis

The basic outline of this thesis is given as follows:

Chapter 1 majorly describes acoustic correlates of vowels, followed with a minor discussion on production of nasal consonants.

Chapter 2 gives the detailed explanation on the methods used for the formant bandwidth extraction of oral and nasal tracts. The methods such as zero frequency filtering, zero time windowing, Hilbert envelope of numerator group delay are discussed.

Chapter 3 mentions the proposed method and displays the results obtained for the extraction of formant bandwidth in oral and nasal tracts.

In **Chapter 4**, the speech performance among different categories of genders is compared with VSA.

Chapter 5 throws light on the role of excitation source and vocal tract system in the perception of vowels.

In **Chapter 6**, the summary and conclusions of the study are presented.

Chapter 2

Methods Used in Formant Bandwidth Extraction

2.1 Introduction

The description of production of vowel sounds and acoustic correlates is explained in detail in Chapter 1. Also, a minor discussion is carried out on the production of nasals and its relevance to resonances of nasal tract in the previous chapter. Incorporating one of the important acoustic correlates for both vowel and nasal sounds, i.e., formant bandwidth as the key factor, this chapter aims to discuss about methods used Chapter 3 for formant bandwidth extraction.

A study of the relative change in the bandwidth across the oral and nasal segments can be made by identifying the spectral resonances within the glottal closed phase. These resonances experience a higher decay during the open glottal phase due to adduction of subglottal and supraglottal tracts. The impulse like excitation in the vibrating source leads to the closed phase of the glottis as the vocal folds shut abruptly. It is during this phase when the subglottal and supraglottal regions get decoupled [26] and as a result, the effective length of the vocal tract gets reduced thereby having the resonances due to the supraglottal part only. This change in the length of the vocal tract induces a change in dominant resonances of the spectrum. Extracting the resonance frequencies and their associated bandwidths accurately is difficult as these keep varying due to the changing vocal tract shapes not only across pitch periods but even within a pitch period (i.e from closed phase of glottis to open phase). Hence, the estimation of resonance bandwidths has to be carefully carried out for short segments of speech (less than one pitch period). Also, it becomes important to keep the size of the analysis window as small as possible to study these changes [36]. Extraction of these bandwidths finds applications in areas such as speech synthesis, acoustics of speech, modeling the physical dimensions of vocal tract configuration, etc.

In the past, model and non-model based approaches were proposed to extract the formant bandwidths. When the speech spectrum is decomposed into amplitude and phase components, then the prominent resonance locations, with a bandwidth associated with these, are known as formants. [37] proved that formant bandwidths obtained with short–time processing can be

approximated to the instantaneous bandwidth of each of the formant. Besides using amplitude component, the instantaneous frequency can also be used to extract formant bandwidths. This was shown in [38] where the formant bandwidths were determined by decomposing the speech signal by passing it through a bank of Gabor bandpass filters and then demodulating each band to get amplitude envelope (AM) and instantaneous frequency (FM) signals. The bandwidths for the formants are then extracted from these instantaneous frequency signals that were separated using energy separation algorithm (ESA). Formant bandwidths were also extracted using an exponentially weighted autoregressive (EWAR) spectral model [39].

Apart from these methods, autoregressive techniques such as linear prediction (LP) analysis (model–based) have also been used to extract formant frequencies and formant bandwidths in speech. One such example is [40] where the formant bandwidths are derived using the phase slope of the z–transform at the poles, which are derived from the peaks of the second derivative of the log–magnitude spectrum (SDLMS). [41] used clustered line–spectrum modeling (CLSM) based method to decompose the speech signal into its dominant frequency components to represent the modal resonant responses of the vocal tract. The modal bandwidths are then derived from the decaying constants. To analyze sub-segmental components of speech, methods like zero-time windowing (ZTW) are used that can reliably highlight the spectral characteristics of short segments in speech [42].

In addition to the above mentioned techniques, group delay function (GDF) has also been used for the extraction of formant bandwidths. Group delay is defined as the negative derivative of the phase response of a resonator [43], [44]. [44] uses GDF to determine the bandwidth for the minimum phase system modeled using a single resonator. The advantages of using GDF for the extraction of formant bandwidths are that the GDF of one resonator will have very less influence on the neighbouring resonators and also the GDF of cascade of resonators is additive [43, 45], unlike the magnitude response which is multiplicative in nature. The GDF is proportional to the square of magnitude response in the neighborhood of resonant frequencies [45]. The important property of GDF of having negligible effect of one resonant peak on the other proved it to be an accurate estimation for extraction of formant bandwidths [45].

Studies in past have used ZTW to detect glottal opening phase, to study dynamics of speech production such as studying glottal activity and characteristics of vocal tract system, in identification of different sounds, to study effect of nasalization on vowels, in estimation of pitch. Study in [46] incorporated ZTW to extract dominant resonances, which give GCI (glottal closure instant) locations that yeild pitch values. Another study [47] used the zero time windowing technique to detect glottal open phase by observing the changes in the vocal tract system. This was conducted by studying the transition from glottal closed to open regions owing to changes in vocal tract system. An attempt was made in [48] to identify turbulent, noise-like exhibiting sounds, fricatives with the help of DRFs obtained using ZTW. This study resulted in high and accurate identification rate, thereby highlighting the significance of ZTW. Also, ZTW has been

used in [26] to distinguish nasal and approximants. When the oral and nasal tracts get coupled to each other, then vowel nasalization takes place. The extent of this acoustic coupling decides the degree to which the vowels get nasalized. Another distinctive feature of using ZTW is that it can be used to study this nasalized vowel region. The work in [49] studied vowel nasalization besides carrying out studies on glottal open regions, studies on changes in DRF spectra wrt vocal tract changes, by employing ZTW analysis. This study has proved that low frequency resonances in nasalized vowels can be highlighted with the help of glottal open regions. Also, a detailed study on vowel nasalization by computing instantaneous spectra using ZTW method was carried out in [50].

This thesis work aims at extracting the formant bandwidths using the Hilbert envelope of the numerator group delay (HNGD) function. The relation between the bandwidth and the GDF is already given in [45]. In this study, we extract the bandwidth at the dominant resonance frequency (DRF) in spectral response derived from short segments in speech, to highlight the change in bandwidth within vowel and nasal segments. The DRF contour is derived using zero time windowing (ZTW) method gives the closed phase region in speech.

This chapter is organized as follows: Section 2.2 explains zero frequency filtering in detail. Zero Time Windowing, Group Delay Function, Hilbert Envelope of Numerator Group Delay and Dominant Resonance Frequency are explained in detail in Sec. 2.3. Section 2.4 summarizes this chapter.

2.2 Zero Frequency Filtering

Speech is the result of time varying vocal tract system excited with impulse like excitation due to vibrating vocal folds. During the production of voiced speech, the vocal folds vibrate and the rate of closure of the vocal folds during each glottal cycle determines the strength of excitation. In the past, popular method such as LP analysis has been used to highlight the characteristics of excitation source [51], [52], [53]. The noise like characteristics present in LP residual results in large and small energy regions. The closing phase of each glottal cycle mostly corresponds to the large energy regions.

The discontinuities caused due to impulse like excitation spreads across zero frequency [54], [55]. Determining the strength of excitation during each glottal cycle is a challenging task. Hence a sophisticated filter must be incorporated to retrieve information about such discontinuities at zero frequency such that it dampens the higher frequencies [6]. Therefore, a zero frequency filter was proposed by [6] to highlight the discontinuities resulted from impulse-like excitation. The filter used was a 2nd order infinite impulse response (IIR) filter having a pair of real poles on the unit circle. [6] obtained a roll-off of 24 dB per octave on using a cascade of two such ideal zero-frequency resonators. This is because, when a speech signal is passed twice through such a resonator, it produces an output that grows/decays as a polynomial function of

time.

The steps involved in zero frequency filtering the speech signal [55], [54] are given below:

(a) Firstly, the speech signal $s\,[n]$ is differenced to remove any DC component present due to the recording device.

$$x\,[n] = s\,[n] - s\,[n-1] \tag{2.1}$$

(b) Now, the differenced speech signal $x\,[n]$ is passed through a cascade of two ideal zero-frequency resonators given by

$$y_o[n] = -\sum_{k=1}^{4} a_k y_o[n-k] + x\,[n] \tag{2.2}$$

where $a_1 = -4, \quad a_2 = 6, \quad a_3 = -4, \quad \text{and} \quad a_4 = 1$

(c) The average pitch period is determined using the autocorrelation of speech segments.

(d) The final step involves trend removal in $y_o[n]$, that is performed by subtracting the local mean computed at each sample.

The resulting signal is given by

$$y\,[n] = y_o[n] - \tfrac{1}{2N+1} \sum_{m=-N}^{N} y_o(n+m) \tag{2.3}$$

where $2N + 1$ is the number of samples in the window used for mean subtraction The signal $y\,[n]$ is the zero frequency filtered signal.

The above mentioned steps can be illustrated with the help of Figs 2.1 and 2.2 [6]. The output obtained after the filtering process of the speech segment shown in Fig. 2.1(a), is shown in Fig. 2.1(b). Zero-frequency filtering causes discontinuities due to impulse -like excitation, that can be overridden by a large DC offset. The mean subtracted signal obtained after performing local subtraction (with window size about one to two times the average pitch period) is shown in 2.1(c) for the filtered output shown in Fig. 2.1(b). This mean subtracted signal is the zero-frequency filtered signal. Sharper zero crossings around epoch locations are seen in the zero frequency filtered signal. The zero crossings can be positive or neagtive according to the polarity of the signal (caused due to recording devices). Fig. 2.1 exhibits positive zero crossings that are much sharper than the negative ones thereby indicating epoch locations. These positive zero crossings coincide with the peaks of differenced electro-glotto-graph (DEGG) signal shown in Fig. 2.1(d). Determination of the polarity of the sharper zero crossings can be carried out

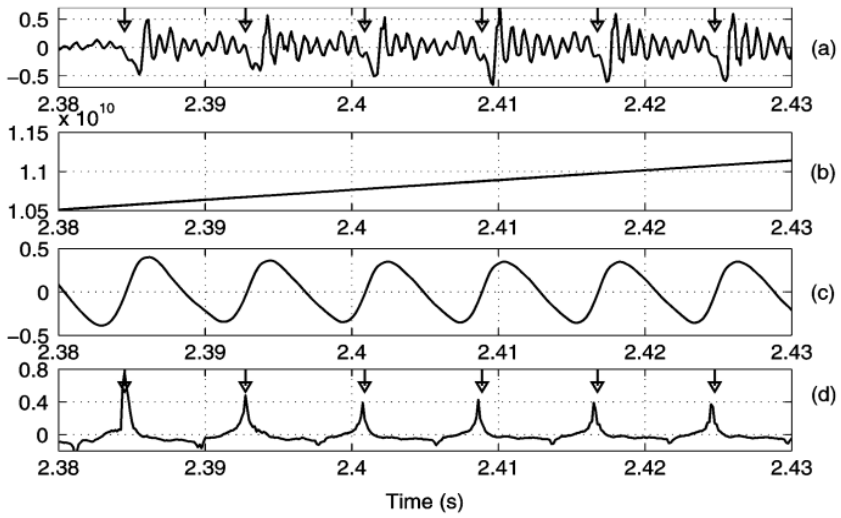

Figure 2.1: Determination of epoch extraction from speech signal [6]. (a) Speech segment taken from continuous speech. (b) The resulting output obtained from cascade of two ideal zero-frequency resonators. (c) Signal obtained from mean subtraction. (d) DEGG signal. The arrows in (a) and (d) indicate the detected epoch locations.

through a comparison of slopes of the filtered signal around both the positive and negative zero crossings over the entire duration of the speech signal.

As, we know that the discontinuities due to impulse is spread throughout the frequency range, which further quests the need to determine the strength of impulse at zero frequency. This can be derived from zero-frequency resonator. And the slope of the zero-frequency filtered signal gives the strength of excitation. A sequence of randomly spaced impulses with their arbitrary strengths are shown in Fig. 2.2(a). This, signal is passed through a ZFF resonator, which computes the zero-frequency filtered signal with sharper zero crossings exactly at impulse locations as shown in Fig. 2.2(b). Fig. 2.2(c) shows the slopes of the zero crossings of the filtered signal that are again proportional to the impulse strengths.

2.3 Zero Time Windowing

Extraction of vocal tract system features is a challenging task as the vocal tract keeps changing its shape continuously with time. Therefore, short-time spectral analysis helps here, but the critical information of the vocal tract system around the glottal closure instant (GCI) may be smeared by the averaging effect either in time domain or frequency domain [42]. Even the size of the window becomes important. In LP based method [56], the inverse filter extracts the vocal tract characteristics, in which the priority lies in the selection of the window size and the order of the prediction. The choice of window size must neither be small nor large. A larger window

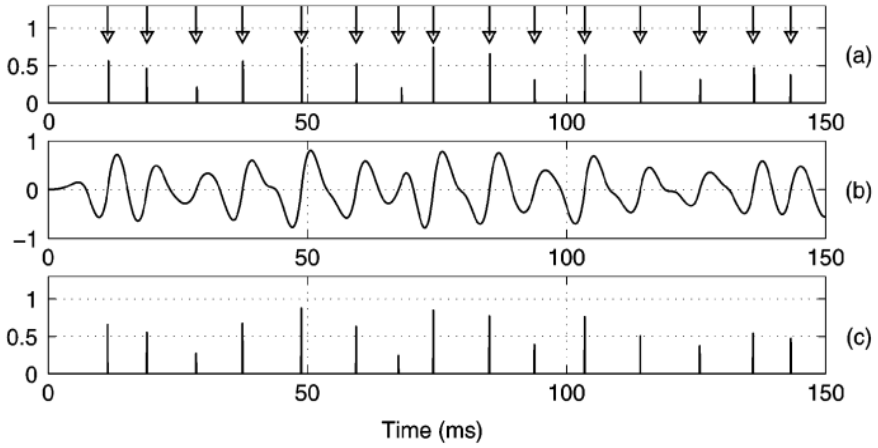

Figure 2.2: The figure is taken from [6]. (a) Sequence of randomly spaced impulses. (b) Zero-frequency filtered signal. (c) Signal showing slope around positive zero crossings

size would result in pitch period affecting the estimated LPCs. Moreover, if the window size is small, then also LPCs get affected due to poor estimation of autocorrelation coefficients from short segments of speech. Hence ZTW method for smaller segments ($\sim 3-5$ ms) of speech is used [47].

In ZTW, the short segment of speech signal is multiplied with a windowing function, which results in an impulse-like signal, resulting in emphasis being focussed on the energy components at the beginning of the window, i.e zero-time. Using ZTW becomes essential as the need for determination of vocal tract characteristics at epochs is felt. As we have seen that the ZFF proposed in [55] , [6], helps in determining the excitation source features such as epochs and instantaneous fundamental frequency, in the frequency domain. But if the desired features are to be highlighted in the time domain, then ZTW operation can be used. ZTW in time domain is analogous to the ZFF method in frequency domain. The ZTW method exploits the additive property of the group-delay [57], [58], [36] function, besides using high resolution, to derive spectral features. ZTW method also helps in extraction of features around starting of epoch, which is again similar to ZFF method emphasizing around zeroth frequency. No effect of the duration of the pitch period is seen on the resulting spectral information.

2.3.1 Basic Steps in ZTW

The basic steps involved in ZTW are given below:

(a) The differenced signal at the sampling frequency of f_s Hz is $s\,[n]$.

(b) The signal having M samples, starting from an arbitrary reference set at $n = 0$, i.e., $s\,[n]$ is defined for $n = 1/40, 1, ..., M-1$.

(c) the DFT length $N \gg M$ is chosen such that sufficient sampling is obtained in frequency domain. The length of the signal $s\,[n]$ is made equal to N by appending $s\,[n]$ with desired number of zeros.

(d) The windowed signal $x\,[n] = s\,[n]\,w_1[n]$, for $n = 0, 1, ..., N-1$ is computed, where $w_1[n]$ is given by:

$$\omega_1[n] = \begin{cases} 0, & n = 0 \\ 1/(4\sin^2(\pi n/(2N))), & n = 1, 2, ...N-1, \end{cases} \tag{2.4}$$

(e) Now, the NGD function of $x\,[n]$ computed, is given by

$$g\,[k] = X_R\,[k]\,Y_R\,[k] + X_1\,[k]\,Y_1\,[k]\,, k = 0, 1, ..., N-1 \tag{2.5}$$

where $X\,[k] = X_R\,[k] + jX_1\,[k]$ is the $N-$point DFT of the sequence $x\,[n]$, and $Y\,[k] = Y_R\,[k] + jY_1\,[k]$ is the $N-$point DFT of the sequence $y\,[n] = nx\,[n]$.

2.3.2 Group Delay Function

Group delay (GD) function is defined as the negative derivative of the phase function. The group delay spectra is additive in nature. The phase function of discrete time signals experiences discontinuities at $\pm\pi$, due to the wrapping effect. Computation of the GD function directly using the discrete-time Fourier transform (DTFT) ($X(\omega)$) derived from the corresponding segment ($x[n]$) helps to overcome this limitation [43].

The DTFT is given by

$$X(\omega) = |X(\omega)|e^{j\theta(\omega)} = \mathcal{F}\{x[n]\}, \tag{2.6}$$

where $\mathcal{F}$ represents the DTFT operation. The GD function from $X(\omega)$ is derived with following steps:

$$logX(\omega) = log|X(\omega)| + j\theta(\omega) \tag{2.7}$$

$$\frac{d}{d\omega}logX(\omega) = \frac{d}{d\omega}|logX(\omega)| + j\frac{d}{d\omega}\theta(\omega) \tag{2.8}$$

$$\frac{X'(\omega)}{X(\omega)} = \frac{d}{d\omega}|logX(\omega)| + j\theta'(\omega) \tag{2.9}$$

By the definition, GD function is given as $g(\omega) = -\theta'(\omega)$. Substituting this in Eqn. 2.9 gives,

$$g(\omega) = -\,\mathrm{Im}(X'(\omega)/X(\omega)). \tag{2.10}$$

To obtain a closed form solution, lets use the Fourier definition of the phase, given by $\theta(\omega) = tan^{-1}(X_i(\omega)/X_r(\omega))$, where $X_r(\omega)$ and $X_i(\omega)$ are the real and imaginary parts of the complex DTFT $X(\omega)$.

$$\theta'(\omega) = \frac{d}{d\omega}\{tan^{-1}(X_i(\omega)/X_r(\omega))\} \tag{2.11}$$

$$= \frac{X_r^2(\omega)}{X_r^2(\omega) + X_i^2(\omega)} * \frac{X_r(\omega)X_i'(\omega) - X_r'(\omega)X_i(\omega)}{X_r^2(\omega)} \tag{2.12}$$

This leads to the following relation,

$$g(\omega) = -\theta'(\omega) = \frac{X_r'(\omega)X_i(\omega) - X_r(\omega)X_i'(\omega)}{X_r^2(\omega) + X_i^2(\omega)}, \tag{2.13}$$

where $X'(\omega) = X_r'(\omega) + jX_i'(\omega)$ is the derivative of $X(\omega)$, and is obtained as the DTFT of $nx[n]$.

As also explained in [45], for a discrete-time signal modeled as an output of a single-pole system, the group delay function is given by,

$$\tau(\omega) = \frac{r_0^2 - r_0\cos(\omega - \omega_0)}{1 + r_0^2 - 2r_0\cos(\omega - \omega_0)}, \tag{2.14}$$

when the pole is located at $z = r_0 e^{j\omega_0}$. The GD when evaluated at the pole location gives

$$\tau(\omega_0) = \frac{r_0}{1 - r_0}. \tag{2.15}$$

The analytical expression for $g(\omega)$ (Eqn. 2.16) when equated with the expression of $\tau(\omega_0)$ obtained from Eqn. 2.15 gives the relation between the radial distance of the resonance pole (r_0), and the group delay expression as,

$$r_0 = \frac{g(\omega_0)}{1 + g(\omega_0)}. \tag{2.16}$$

The bandwidth of such a resonance pole is related to its radial distance by the relation,

$$r_0 \approx e^{-\pi B T_s}, \tag{2.17}$$

where B is the bandwidth of the resonant pole, and T_s is the inverse of sampling frequency (f_s). Eq. 2.16 and Eq. 2.17 give the relation between bandwidth of a resonance and the group delay expression, as follows,

$$B = -\left(\frac{f_s}{\pi}\right) log\left(\frac{g(\omega_0)}{1 + g(\omega_0)}\right) \tag{2.18}$$

The relation holds for multiple pair of resonant poles, although the contribution of the presence of multiple resonances in the GD response to a resonance in consideration is even less than 10% [45].

2.3.3 Hilbert Envelope of Numerator Group Delay

The present study uses the numerator of group delay (NGD) function to represent the spectral resonance. Zero time windowing method is used to compute the instantaneous spectra of the segments in speech.

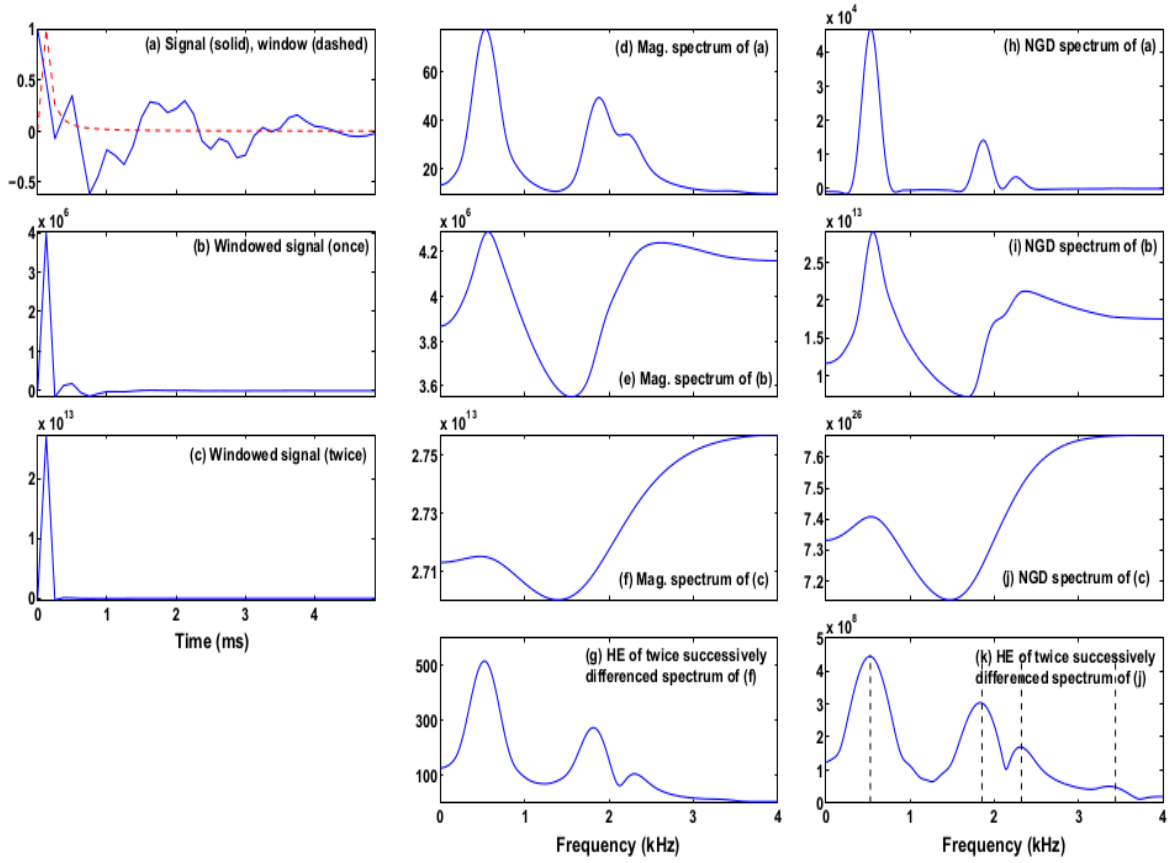

Figure 2.3: Illustration of ZTW and HNGD operations of a voiced speech segment (a) Signal and the window function are indicated by solid and dashed lines respectively.(b) Windowed signal (once) (c) Windowed signal (twice) (d) Magnitude spectrum of (a). (e) Magnitude spectrum of (b). (f) Magnitude spectrum of (c). (g) HE of twice successively differenced spectrum of (f). (h) NGD spectrum of (a). (i) NGD spectrum of (b). (j) NGD spectrum of (c). (k) HE of twice successively differenced spectrum of (j). True formant frequencies determined are indicated by dashed vertical lines.

The method uses a heavily decaying window to segment the signals. The spectral response for small segments is computed using the Hilbert envelope of the numerator of the group delay (HNGD) function.

Figure 2.3 illustrating ZTW and HNGD operations is taken from [42]. Fig. 2.3(a) shows speech segment with an arbitrary epoch. Fig. 2.3(b) shows the windowed signal obtained after

performing the windowing operation by $w_1[n]$ once. The windowing operation is performed twice in Fig. 2.3(c). Figs. 2.3(d), 2.3(e) and 2.3(f) illustrate the computed DFT spectra of Figs. 2.3(a), 2.3(b) and 2.3(c) respectively. The spectra of windowed signals in Figs 2.3(e) and 2.3(f) are the smoothed versions of the spectrum in Fig. 2.3(d). The formants lost due to this smoothing operation in Fig. 2.3(f) are retrieved by computing the HE after twice successively differencing the plot in Fig. 2.3(f). The resulting formants can now be seen in Fig. 2.3(g).
The spectral resonances can be better highlighted through the numerator of the group delay function. So, Figs. 2.3(h), 2.3(i) and 2.3(j) represent the NGD plots of the plots shown in Figs. 2.3(a), 2.3(b) and 2.3(c) respectively. Now, HE of NGD is computed and the resulting plot indicating spectral peaks given by dashed vertical lines are shown in Fig. 2.3(k). It can be observed that Fig. 2.3(g) shows only three formant peaks whereas Fig. 2.3(k) shows even the fourth spectral peak. This is due to the additive and multiplicative property of the individual resonances in the magnitude and phase spectra respectively [57].

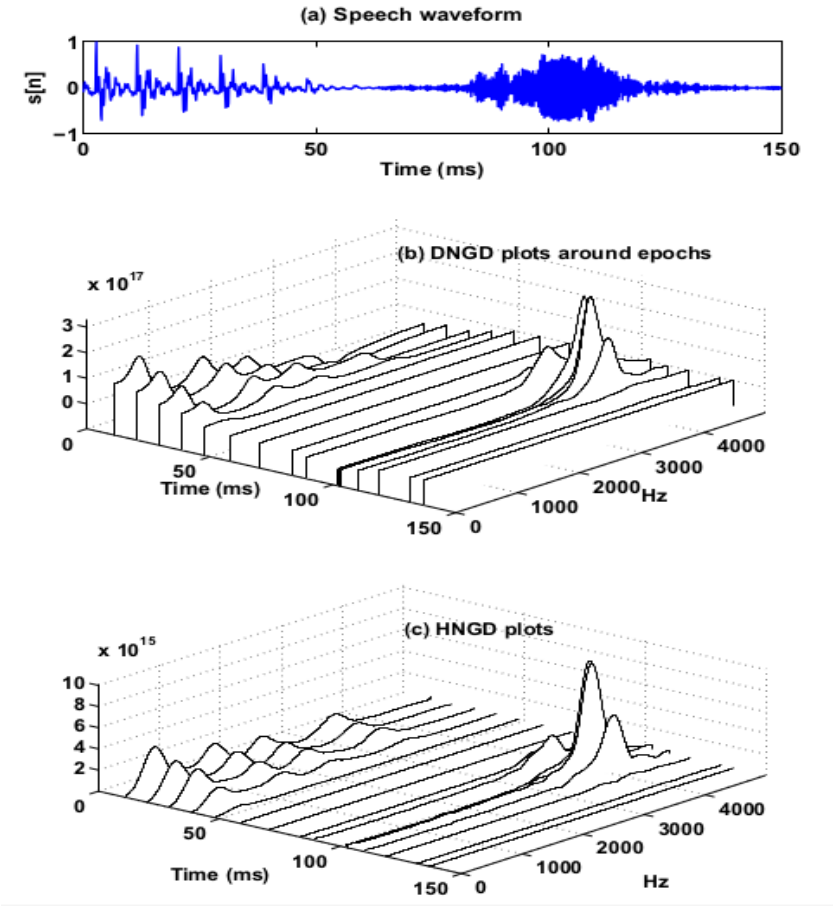

Figure 2.4: Selection of DNGD plots around epoch. (a)Speech waveform. (b) DNGD plots selected around epoch locations. (c)HNGD plots selected around epoch locations.

Figure 2.4(b) and (c) portrays DNGD and HNGD plots respectively in 3D representation for the speech segment shown in Fig. 2.4(a).

2.3.4 Dominant Resonance Frequency

The HNGD spectrum displays peaks corresponding to resonances of the vocal tract system. The peak having the highest amplitude arises due to the effective length of the vocal tract system as explained in Sec. 2.1. This peak is the strongest one and therefore known as Dominant Resonance Frequency (DRF) [59]. As, we know, during the closed phase of the glottis, the effective length of vocal tract decreases, which in turn raises the DRFs, when compared over the open phase of the glottis.

2.4 Summary

This chapter highlighted the methods used for extraction of formant bandwidths for oral and nasal tracts implemented in Chapter 3. As the experiment implemented in Chapter 3, carries out analysis for closed phase of the glottis over a short segment of speech, it becomes important to extract features around epoch locations. Therefore, methods such as ZTW and ZFF that focus their filtering operations around epoch locations were discussed elaborately. Also, the exploitation of the additive property of the group delay function to extract HNGD spectra containing DRFs has also been discussed.

.

Chapter 3

Study of Closed Phase Resonance Bandwidths For Oral and Nasal Tracts

3.1 Proposed Method

The ZTW method (discussed in Chapter 2) is a reliable method to derive the spectral characteristics for smaller segments ($\sim 3-5$ ms) in speech. The HNGD is obtained at every sample of speech, which gives a good temporal resolution. A small analysis window however, imparts more bandwidth in frequency domain. The group delay function has been proven to exhibit sharper peaks at the spectral resonance locations, in comparison to short time Fourier spectra. The dominant resonance contour obtained using ZTW analysis with smaller analysis window helps to identify the glottal closed and open regions in speech.

For the present study, we use the smaller analysis window duration ($\leq$ average pitch duration of signal) to derive the prominent resonances during the glottal closed phase, for the vowel (oral) and nasal regions. The study hypothesizes that the prominent resonances obtained in the glottal closed phase region in nasal segments will exhibit larger bandwidths as compared to those obtained during the oral segment. This is due to lossy behavior of the nasal tract, which is captured during the glottal closed phase. During the closed phase, the vocal tract system response is attributed to the supra-segmental tract, hence the change in bandwidth is in proportion to the adduction of nasal cavity to the oral cavity. The study further hypothesizes that the change in bandwidth is independent of the VC pair in consideration.

The supraglottal cavity appears as open oral tract in case of the production of vowels, extending from the closed glottis region till the lips. This cavity is relatively longer in the case of production of nasal segments due to the adduction of a longer nasal tract, and closure of the oral tract. The nasal cavity exhibits higher impedance and hence larger value of decay as compared to the oral cavity. Closure of the oral cavity further introduces a zero in the nasal spectra in the 1 kHz frequency range. A cumulative effect of an increment in resonance bandwidth due to increased decay, and the presence of zero in the spectra in vicinity of the resonance has to be taken into consideration during the study of resonance bandwidth. Therefore the study focuses

on highlighting the relative increment in bandwidth between the vowel and nasal segments, while identifying the absolute values can be attempted as a further exercise.
The method uses Eq. 2.18 to derive bandwidth values for resonances obtained during the glottal closed phase, for nasal and vowel segments. The resonance locations and their respective group delay values are obtained from the HNGD spectra obtained using the ZTW method. A ZTW analysis is performed using a window duration of 4 ms. The following Fig. 3.1 shows the dominant resonances and their respective bandwidths obtained using the proposed method.

This chapter is organized as follows: Experimental study is described in Sec. 3.2. Section 3.3 discusses the results obtained. Summary and conclusions are discussed in Sec. 3.4.

3.2 Experimental Study

The utterances from TIMIT database, recorded at a sampling rate of 16 kHz are used to study the relative difference in the impedance of the oral and nasal cavities, based on the change in bandwidth of the respective dominant resonances. The effect of multiple poles can be countered by computing the group delay at lesser dominant poles and eliminate it from the delay occurring at the dominant pole in concern. The experiments study the change in bandwidth and hence the impedance characteristic in the supraglottal oral and the nasal cavities. The speech segments for 20 speakers, 10 female and 10 male, corresponding to different vowel–nasal pairs are chosen. The present study is independent of the choice of CV/VC pairs as it is an acoustic characteristics based study, and relies on parameters derived from short segments. The glottal closed phase duration within these segments is identified using a method proposed based on dominant resonances in the HNGD spectrum. These segments consist of VC (V-vowel, C-nasal) or CV clusters. A total of 90 segments of various combinations of VC/CV clusters are used where the vowel set comprises of all except the close back vowels, and the nasal set include /m/ and /n/. The close back vowels are not considered due to the fact that their dominant resonances are closer to that of nasals, hence cannot be resolved. The boundaries for CV/VC clusters were selected using the transcription files provided by the TIMIT database. The corresponding resonance bandwidths were plotted for the respective segments. The respective bandwidths were then compared for the nasal and oral clusters.

Fig. 3.1(a) shows the speech signal corresponding to the vowel-nasal transition segment. The glottal closure instants (GCIs) derived using the zero frequency filtering method are also marked (in red) in Fig. 3.1(a). Fig. 3.1(b) shows the dominant resonance contour obtained from the HNGD spectra. The HNGD spectra has been obtained with a ZTW analysis using a window duration of 4 ms. The HNGD spectrum is obtained at each sampling instant and the dominant resonance is derived as the strongest peak in the spectrum. Fig. 3.1(c) shows the strength of the dominant resonances which reflect the corresponding $g(\omega)$ values at these resonances. Fig. 3.1(d) shows the bandwidth of the resonances obtained using the proposed method.

3.3 Results

The bandwidth values obtained for dominant resonances for vowel and nasal segments have been illustrated in Fig. 3.1(d). Following this, the bandwidth values for 90 different segments pertaining to 20 speakers (10 male and 10 females chosen randomly), are plotted for vowel and nasal segments.

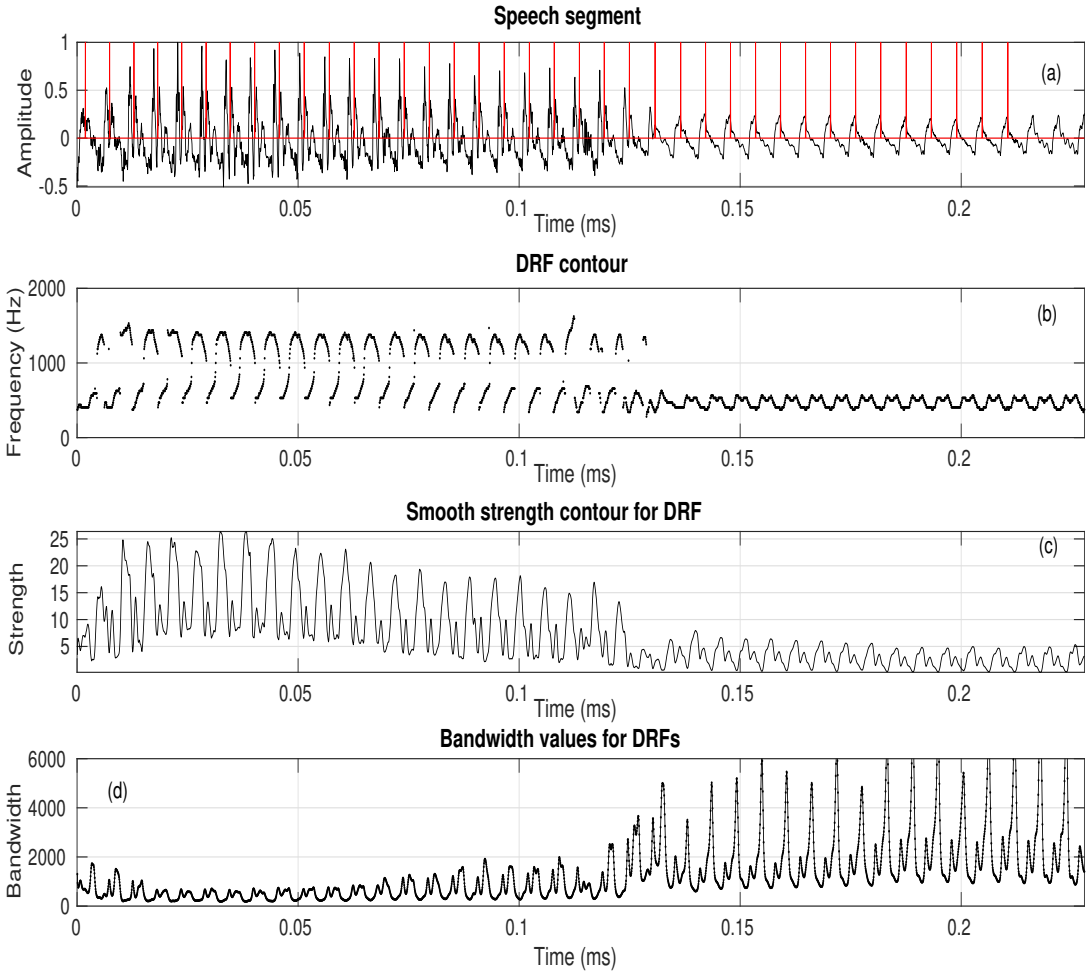

Figure 3.1: Formant bandwidth obtained for VC segment 'aim', for the prominent resonances in HNGD spectrum with analysis window size of 4 ms. (a) Speech signal with GCI locations. (b) Dominant resonances obtained using HNGD spectra. (c) Strength of resonances. (d) Bandwidth of the resonances.

These values are obtained from glottal closed phase regions. Bandwidth values are normalized with respect to the highest value, and plotted as histogram curves for each of the 20 speakers. Fig. 3.2 shows the histogram curve for the closed phase resonance bandwidth for all 90 vowel and nasal segments. As can be seen from Fig. 3.2, the nasal segments exhibit an increased bandwidth compared to vowels and a result nasals exhibit large impedance in comparison to

vowels. This reflects a higher impedance for the nasal tract.

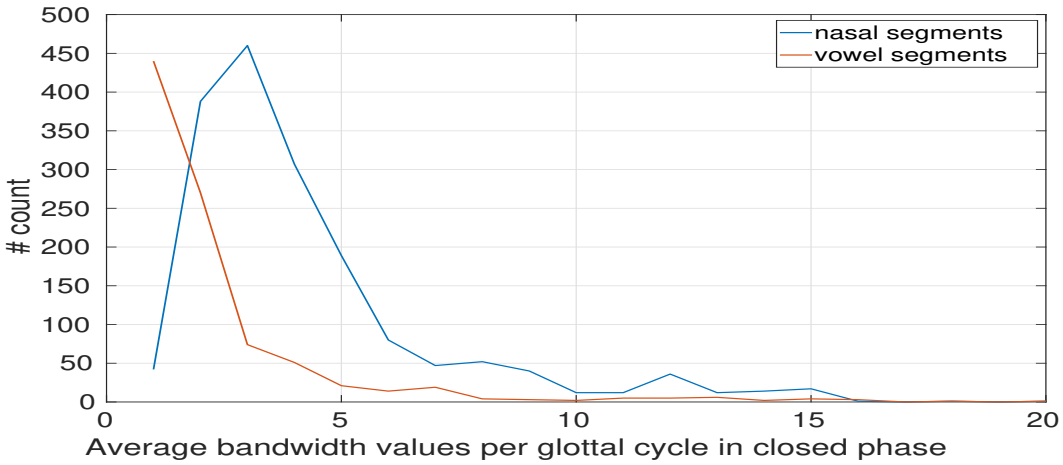

Figure 3.2: Histogram curve representing average bandwidth values across closed phase for vowel and nasal segments in TIMIT.

3.4 Summary and conclusions

In this work, a relative comparison of closed phase resonance bandwidths for nasal and oral tracts is made. The closed phase dominant resonance obtained using a short-time processing based on the ZTW method is considered for this study. The bandwidth values are obtained using the group delay based formulation, and are computed for the closed phase DRFs. These values are compared for pairs of nasal and vowel regions occurring in VC/CV pairs. The results show that nasals have higher bandwidth, owing to the higher impedance in the nasal tract than the vowels. The study can further be extended to derive regions of nasalization in vowels. It can also be improved to determine actual bandwidth of oral and nasal cavities, by considering multiple resonance model to derive bandwidths.

Chapter 4

Comparison of Speech Performance Among Men, Women and Children Using Vowel Space

Speech intelligibility depends on how well vowel articulations take place. Much emphasis is put on vowels, as they form the nucleus of a syllable. It becomes very difficult to recognise a speech segment that includes only the consonant sounds. Thus vowel articulatory space, or to be precise, Vowel Space Area (VSA) serves to exploit a speaker's articulatory space. Moreover, as already mentioned in Section 1, VSA is an important acoustic correlate of vowels. This is because it very well portrays the size of the vowel articulatory working space [18] and determines the accuracy of vowel articulations [60]. This not only is associated with speech production but also relates to speech perception, as the compression/expansion of VSA is directly proportional to the listener's identification [60]. Hence, VSA can be exploited at acoustic-perceptual level too.

The speaker's speech performance is influenced by various factors like vocal tract configurations, articulatory movements, shape of articulators, and various other factors such as gender, age, rate of change of vocal tract, phonetic context and rate of speech production [61]. As we know from Chapter 1, vowels are produced with a relatively open vocal tract [62] and hence display a clear spectral (formant) structure caused due to their continuous, quasiperiodic vibrations of vocal folds without any obstruction [63]. This chapter uses two acoustic correlates viz., formants and VSA (discussed in Chapter 1) to assess speech performance by analyzing vowel articulations among different categories of gender.

The vowel articulations can be well displayed by vowel space (VS) [64]. Assessment of speech performance using VS would help in analyzing the articulatory-acoustic relationship in speech sounds. VS gives insight on how acoustic aspects of speech such as formants keep changing with age across genders (men, women and children) [65]. The area of VS has been used as an effective metric to visualize speech performance [64] and assess speech intelligibility [66], [67]. It is the space (quadrilateral or triangular) bounded by the lines joining F1-F2 plane. This area defines the space where the speaker spends most of his time articulating the vowel sounds. Thus, the boundaries of this area marks the extreme positions (cardinal) that the tongue can

take in producing them. The corner vowels define the periphery of the vowel system [68], and hence would enclose the complete region involved in the production of vowels. In an ideal case, the shape of vowel space is quadrilateral. When the articulatory gestures for vowel sounds reach their target positions completely, then the defects in speech production are minimal, giving rise to the VS that is closer to the ideal one. This results in good speech performance. The expansion of vowel space reflects greater articulation, thereby gaining higher intelligibility [69, 70, 71, 72].

VS has been used in for degrees of articulation like Hyperarticulated speech and Hypoarticulated speech in [64]. Here it was proved that hyperarticulated speech leads to a VS expansion. Also, VS has been used to study speech disorders and speech deficits caused due to cerebral palsy [73], amyotrophic lateral sclerosis (ALS) [68], stuttering [74], dysarthria in ALS [75], partial glossectomy [76], closed head trauma and cerebellar lesions [77], [78]. It has been used to study the speech performance in various neurological conditions like down syndrome [67], parkinson's disease (PD) [79], suicidality [80], depression, post-traumatic stress disorder (PTSD) [81], multiple sclerosis [82] . Some studies like [16] have used other metrics such as Formant Centralization Ratio (FCR), Logarithm Vowel Space Area (LnVSA), F2i/F2u ratio that helped for measuring speech performance in neurologically disordered people. [83] used pentagonal VS for studying disordered speech. Besides these, VSA has also been used in speech perception and production with cochlear implants [84]. Also used to study cross-language comparisons [85].

Other techniques such as FCR, LnVSA, F2i/F2u ratio have also been used in various studies for assessing speech performance in dysarthric speech and speech related to other disorders. FCR (Formant Centralization Ratio) has been proved to serve better than VSA, LnVSA, F2i/F2u as it has maximum sensitivity to vowel centralization and minimum sensitivity to interspeaker variability [16]. When the articulatory movements are reduced, such that the place and degree of vocal tract constriction cannot be fully achieved. then this type of articulatory undershoot results in vowel formant centralization, in which formants having normally high frequencies tend to have low frequencies and formants having normally low frequencies possess achieve higher frequencies [16].

[20] proposed VSA to represent this centralization. Thus, if speech performance is to be improved, then VSA should get expanded [86]. This type of centralization is given by FCR which is expressed as FCR = (F2u + F2a + F1i + F1u)/(F2i + F1a), where F2u is the frequency of the second formant of the vowel /u/ and F1i is the frequency of the first formant of vowel /i/ and so on. FCR is designed in such a way that, it should increase with centralization and decrease with vowel exapnsion [16]. Another metric called LnVSA, a logarithmically scaled version of the VSA, is used to scale the formant frequencies logarithmically [16]. LnVSA was proved to be less sensitive to interspeaker variability. Also, another metric called the F2i/F2u ratio differentiates dysarthric speech of IPD (Idiopathic Parkinson's disease) from normal speech [87], speech in speakers with Down syndrome from normal speech in typical children [88]. As the F2i/F2u ratio itself accounts for two vowels /i/ & /u/, it means, it should contribute for

anterior-posterior movements of the tongue and also the rounding and unrounding of the lips as these movements will affect the F2u & F2i. This means that F2i/F2u ratio should be less for more articulatory undershoot and should be more to indicate improved articulatory movements [16]. Thus, a greater value of F2i/F2u ratio would account for expansion of vowel space area.

We can expect that vowel formants show centralization as FCR, VSA. LnVSA are all based upon the building of vowel centralization [16]. In literature, FCR and F2i/F2u ratio differentiate dysarthric speech from normal one effectively unlike VSA and LnVSA. FCR is proved to be better than F2i/F2u ratio, VSA & LnVSA as it is less sensitive to interspeaker variability. The various factors affecting interspeaker variability may be based upon age, gender, size and shape of the vocal tract [89], [90]. With this, one can say that anatomical and physiological factors attribute to interspeaker variability. Other factors such as severity of dysarthric speech, the nature of speech task, the phonetic environment in which the vowels in VSA are measured [91], [92]. Also, clear speech was found to be less intelligible than conversational when the raising of F2 caused vowel frontness in [70]. But again, VSA was found to be larger for clear speech causing speech to be more intelligible [93] and smaller for read speech [80]. Also, the speaking rate would affect VSA. One such example is [94], where, the faster speaking rates among speakers resulted in compression of vowel space area when compared with slower rates.

VSA has shown poor performance compared to FCR as it is highly sensitive to interspeaker variability [16]. Also, FCR is more effective when compared to F2i/F2u ratio as the latter contains only 2 vowels and 1 formant, whereas, FCR is inclusive of three vowels and two formants [16]. But VSA is constructed taking three or four vowels also, thereby advantageous over FCR and F2i/F2u ratio. But the disadvantage with VSA over FCR is that, it is more sensitive to interspeaker variability. The disadvantage of both VSA and FCR is, both of them rely on the vowel format measurements only at specific instants of time (i.e point(triangle) or corner vowels(quadrilateral) but not during the significant duration (course) of actual speaking [64]. This is the second limitation of VSA.

For the present work, we compare the speech performance among three categories (men, women and children) by constructing vowel space using F1, F2 extracted from LP spectra of four cardinal vowels [i], [æ], [ɒ] and [u]. The fact that vowels can be distinguished from each other based on two classifications, position of the tongue and height of the tongue [25] is explained in detail in Chapter 1. The way F1 and F2 values attain for different vowels as per the vertical height and horizontal length of the tongue is clearly stated in Chapter 1. Exploiting this relation, formants F1 and F2 are computed spectrally from LP spectrum using LP analysis [95], [56]. These computed formants are then used to construct VSA.

The work in this chapter is organized as follows: Section 4.1.1 describes the procedure for extraction of linear prediction (LP) spectrum using LP analysis. Section 4.2 explains about the database used and experimental details. Section 4.3 presents results obtained and related discussion. Conclusions of the study are presented in Section 4.4.

4.1 Proposed Method

As discussed already, the construction of VSA requires formants F1 and F2. The formants are derived from LP spectrum by using LPA [96] discussed in Sec. 4.1.1. Also, Fig. 4.1 illustrates the LP technique. Studies from past that have incorporated speech synthesizers using all-pole filter model to synthesize speech signal show that, such a model is quite adequate for speech production [95], [97], [98]. Extraction of spectral envelope by this method is found to be less prone to the consequences of high pitch on the spectrum [99]. Also, study from [100] reported that the naturalness of the speech signal was affected to a noticeable difference when synthesized through an all-pole model.

4.1.1 Extraction of Formants from Linear Prediction spectra

Linear Prediction (LP) analysis is a widely used method for extraction of excitation source and vocal tract system information from a given speech signal. The dependence of linear prediction on the all-pole model of the vocal tract is well known. For sonorant sounds like vowels, the vocal tract can be modelled as an all-pole filter (whose transfer function consists of only poles) due to only resonances present. These resonances represent the formant frequencies of the vocal tract. So, the work on VSA in this chapter uses LP method realised with an all-pole filter model. The advantage of selecting LP method is attributed to the characteristics of the spectral envelope extracted.

In LP analysis, the present sample of a given speech signal is approximated as a linear weighted sum of its past samples given by the following equation:

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{4.1}$$

where

$\hat{s}(n)$ = modelled present sample
$s(n-k)$ = past samples
p = order of LP
a_k = LP coefficients

Excitation source information known as LP residual is obtained from the following equation

$$E(z) = S(z)(1 + \sum_{k=1}^{p} a_k z^{-k}) \tag{4.2}$$

where

$E(z)$ = z - transform of LP residual
$S(z)$ = z - transform of given speech signal

Vocal tract system is modelled as an all pole model in LP analysis. The spectral information of vocal tract in a given quasi-stationary period can be obtained from the following equation.

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{4.3}$$

Figure 4.1(a) shows a quasi-stationary segment of the vowel /u/. Fig. 4.1(b) and Fig. 4.1(c) represent LP residual and LP spectrum obtained from this segment.
Formants are extracted from LP spectrum by using peak picking algorithm. Formant values extracted from LP spectrum of above speech segment are marked in Fig. 4.1(c).

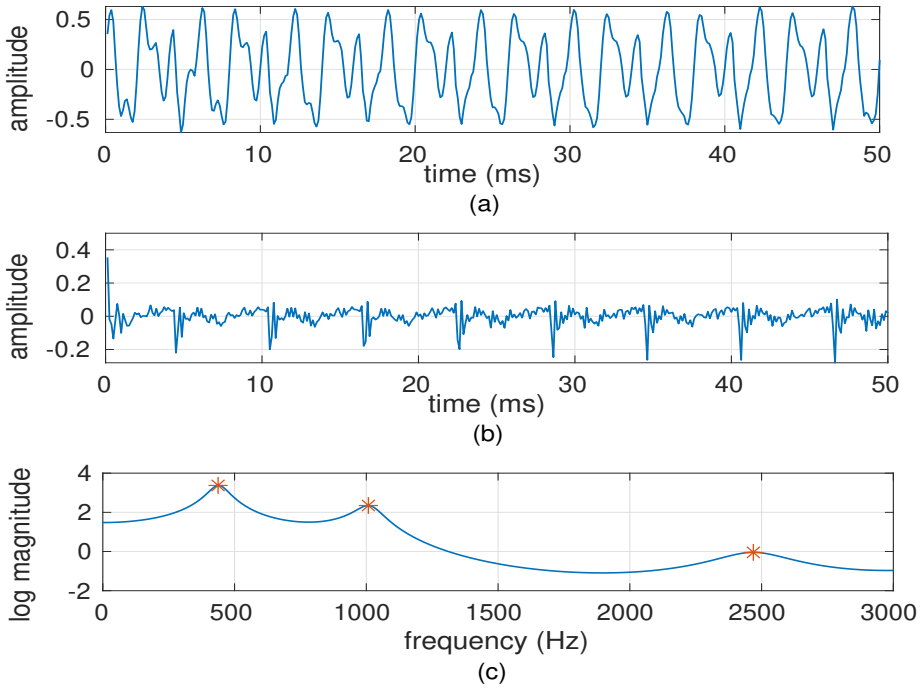

Figure 4.1: (a) Speech signal of vowel [u], (b) LP residual of (a), (c) LP-derived magnitude spectrum of (a)

4.2 Experimental Details

4.2.1 Database

For the current study, Michigan Vowel Dataset (MVD) [15] is used. The dataset contains audio recordings of vowels [i,, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, ɜ, e, o] for 50 males, 50 females, 50 children (25 girls and 25 boys), in h-V-d syllables like 'heed', 'hid', 'head', 'hood', and so on. Majority of

the recordings were of the speakers from southeastern and southwestern parts of the Michigan state, whereas the rest from middle and northern parts. One benefit of using this database arises from the fact that much attention was paid to the dialectal specification of the speaker. This was done mainly to focus on the speakers' production of /ɑ/-/ɔ/ vowels, as this distinction is not maintained by majority of american english speakers. Another advantage from this dataset is the provision of vowel data like vowel duration and F1, F2 values. The vowel steady state region values and formant centre pattern values were provided for vowel duration and formant frequencies respectively.

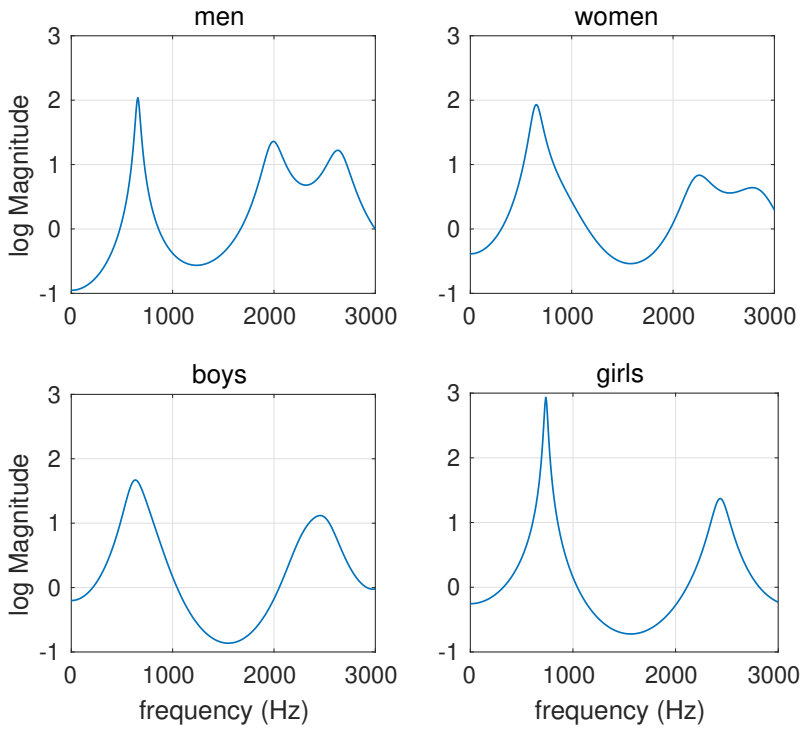

Figure 4.2: LP spectra of vowel [æ] for men, women, boys & girls

4.2.2 Experiment

A 20 millisecond rectangular window is placed at the center of a vowel utterance in order to extract quasi stationary segment. LP spectrum and formants of this segment is derived using method explained in Section 4.1.1. In this manner, formant values are obtained for all corner vowel utterances in MVD. Figure 4.2 illustrates LP spectra of [æ] derived from a single respective sample utterance of men, women, boys and girls. In Fig. 4.4, distribution of formant values in F1-F2 plane for all the above mentioned categories is presented. The LP order was set at p=8 through out the experiment. Sampling rate of the speech recordings was maintained at 16 kHz. The colouring scheme used in Fig. 4.4 is as follows: Red-[i], Green-[æ], Blue-[ɒ] and Black-[u].

X-axis represents F1 ranging from 200 to 1000 Hz and the Y-axis represents F2 ranging from 0 to 4000 Hz. Mean values of F1 and F2 and their respective standard deviations are presented in Table 4.1. Using the mean of F1-F2 values from Table 4.1, the VS for men, women and children was presented.

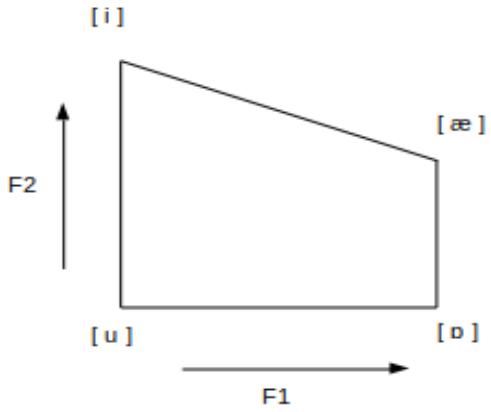

Figure 4.3: Ideal vowel space (VS) for the corner vowels [i], [æ], [ɒ] and [u]

4.3 Results and Discussion

Figure 4.2 reveals F1, F2 and F3 frequencies for a male and female speaker. But for children (boy and girl), there is no F3. This could be due to the inability of the children to develop complete dimensions of vocal tract (such as pharnygeal length, oral cavity length etc,.) till the age of 15 [65]. Hence the acoustic and articulatory characteristics could be affected by the anatomical aspects of a person's vocal tract. This is in agreement with the results shown in [101]. From Fig. 4.2, it can also be noted that F1 is low and F2 is high for all the categories, implying that the vowel [æ] must be a close vowel. Also from Fig. 4.2, it can be noted that none of the formants for the same vowel among three different categories are same. This is due to the different changing vocal tract configurations and also due to the different movements of tongue and lips giving rise to varied formant frequencies.

Ideal articulation (a quadrilateral VS) is shown in Fig. 4.3. Here corner vowels are distinctly apart and occupy four corners in the vowel space forming a quadrilateral, and hence reflecting the amount of time spent by the speaker enclosed by this space.

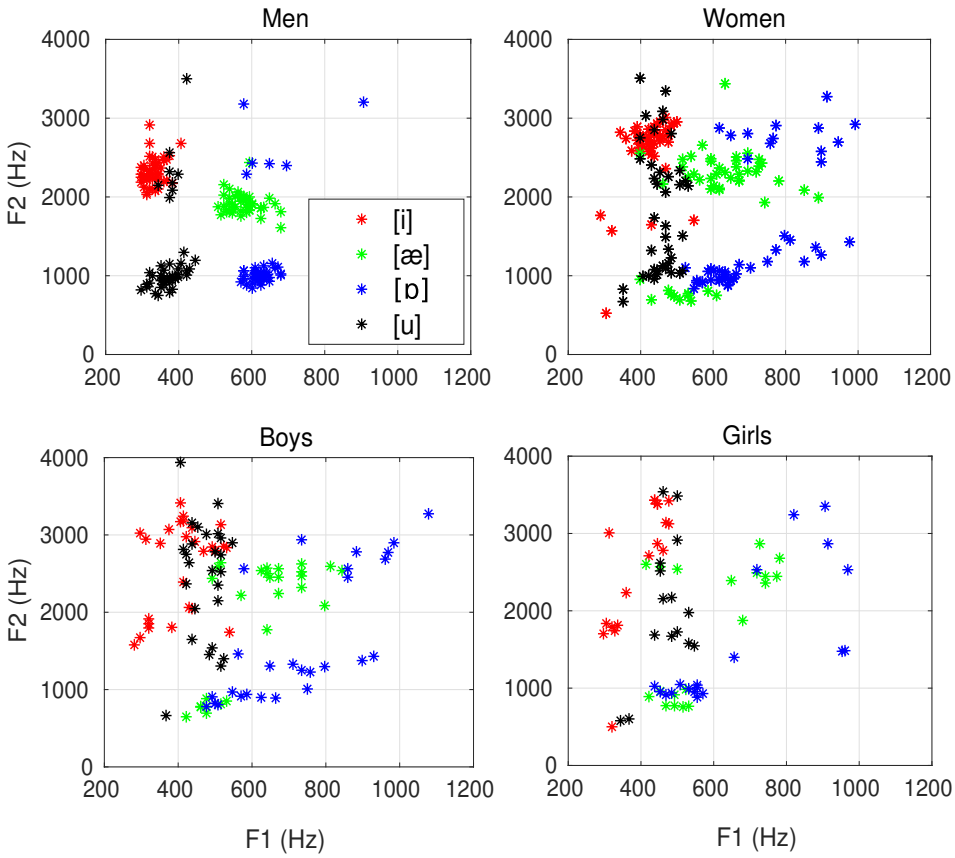

Figure 4.4: F1-F2 distribution of corner vowels for Men, Women and Children (Boys and Girls) for analysis window size of 20 ms

Figure 4.4 shows the distribution of obtained F1, F2 values across all the categories. Here, we are analyzing the articulatory dynamics of the three categories, viz, men, women and children. From Fig. 4.4, we can observe the following: The front-close vowel [i] is clustered well across all categories. It can be observed from Table 4.1 that, [i] has least standard deviation values across all vowels inside an individual category. The back-close vowel [u] is clustered well in men, while in women we notice a slight dispersion in vertical direction. This dispersion is further prominent in children. It can be noted while dispersion in [u] has an increasing trend from men, women and children (men $<$ women $<$ children), this follows the same order w.r.t decrease in vocal tract length and increase in pitch. Owing to smaller vocal tract length and also other dynamics, plots suggest the tongue position in women and children while articulating [u] is sometimes oriented towards middle and front positions of the vocal tract.

For open vowels [æ] & [ɒ], we see a horizontal dispersion in women and boys. Since horizontal dispersion implies wide range of F1, and F1 corresponds to openness of the vocal tract, the dispersion in distribution plots suggest that there is not much distinct variation in closing and

opening of the vocal tract in women and boys, during the articulation of [æ] and [ɒ]. While they are clustered well in the case of men, in the case of girls, we see distribution of [ɒ] has shifted towards [u].

Table 4.1: Mean and Standard deviation values of corner vowels for different categories

Category	**Vowel**	**Mean**		**Standard Deviation**	
		F1	**F2**	**F1**	**F2**
Men	[i]	331.6	2325	24.51	178.2
	[æ]	577.4	1913.5	42.23	130.53
	[ɒ]	628.3	1219.6	51.73	589.72
	[u]	371.8	1233.7	33.71	592.61
Women	[i]	424.6	2619.1	48.5	451.09
	[æ]	607.3	1978.4	106.55	704.13
	[ɒ]	704.9	1527.2	128.20	789.74
	[u]	454.9	1837.3	42.51	803.47
Children (Boys)	[i]	409.4	2593.8	78.01	586.73
	[æ]	611.4	1938.9	122.71	790.15
	[ɒ]	728.3	1649.3	175.76	841.36
	[u]	472.4	2464.5	45.99	751.92
Children (Girls)	[i]	394.1	2572	68.59	842.26
	[æ]	583.1	1794	132.15	847.16
	[ɒ]	661.2	1555.9	194.3	866.68
	[u]	470.3	2051	56.58	874.66

Figure 4.5 shows vowel quadrilaterals for men, women, boys and girls calculated from mean values of F1, F2 shown in Table 4.1. From Fig. 4.5, it can be observed that, women have smaller VS compared to men (which may be because of smaller vocal tract area). The VS of men is nearly similar to that of ideal articulation shown in Fig. 4.3. The vowel space among children has taken shape close to a triangle majorly because [ɒ] is moving towards [u]. The VS of boys seen from Fig. 4.5 may be attributed to the linear relationship between age and formant frequency changes [65] while this may not be true for girls. The VS of boys forms a quadrilateral space but is much narrower and steeper, compared to that of men, women and girls, which could be due to lack of clarity while articulating vowel sounds, as clear speech is one of the factors for increased formant frequency leading to an expanded vowel space [102]. The compression of vowel space in kids may also result from the limitations of the LP method being used as LP method may miss formants which may lead to reduced vowel space. Hence, there is a need to validate the shrinkage of vowel space using other signal processing methods/tools. The shrinking of VS in women, could also be due to the dialectal differences of these speakers. These dialectal differences may change the kinematics of tongue (height, position) and hence

may influence the acoustic dimensions thereby influencing the VS. Another factor that could also account for the this change may be the relation between fundamental frequency (f_0) and the related formant frequencies. Thus, the shrinkage of VS for women may be due to the higher (f_0) causing formant frequencies to decrease, leading to a reduced VS.

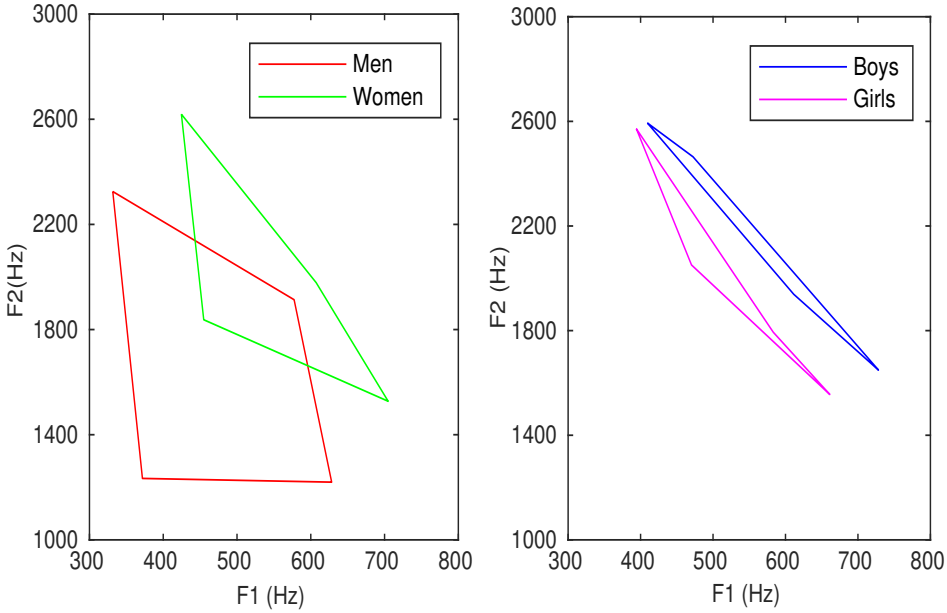

Figure 4.5: Vowel quadrilaterals for Men, Women and Children

4.4 Summary and conclusions

In this study, speech performance of men, women and children is studied. An ideal speech performance is considered to form quadrilateral (trapezoid) in F1-F2 plane. The study suggests while articulation of adults form quadrilateral shape in F1-F2 plane, the VS of children is not distinctly quadrilateral because of lack of clear variations in open-close configuration of vocal tract. The study also suggests that [i] is articulated well by all speakers, and vowels that have more dynamic articulations are [æ] and [ɒ]. Hence, this work shows how the relationship between acoustic and articulatory characteristics of speech influence VS across gender and age.

Chapter 5

Understanding the Role of Excitation Source and Vocal Tract System in Vowel Perception

Speech is produced as a result of pseudo-periodic impulse train originating at the glottal source, filtered by a time varying vocal tract system response. Vowels form the syllable nuclei as they are the voiced sounds produced with open articulation leading to a well-defined spectral structure excited with sharp impulses. This spectral structure is due to the characteristic resonances produced in the vocal tract system and hence vowels are the most significant speech units for both speech production and speech perception [63]. Studying the contribution of excitation source and vocal tract system components may be applied in areas such as speech analysis, speech modification, speech separation, speech synthesis and voice conversion.

Perceptual identity of a vowel is independent of the speaker, the speaker group, the vocalization type, the dynamics and the pitch. A listener's auditory characteristics, language, dialectal background and previous language experience [103] also play a crucial role in the identification of vowels. [104] explained a target model for vowel perception that focussed on three characterizations of vowels viz. articulatory, acoustic and perceptual.

In articulatory terms, the vowel targets are represented by the shapes of the vocal tract involved in the production of vowel sounds. Acoustically, these vowel targets are given as the points represented by the F1/F2 or F1/F2/F3 formants [105], [103]. From perceptual point of view, the first two formant frequencies give the sufficient information for identifying a vowel [106] that was consistent with [107], [108]. It is widely believed that the frequencies of the first two or three lowest formants are the most important for vowel perception [103] and that secondary spectral properties such as formant bandwidth, global spectral tilt, or formant amplitude are relatively unimportant for preserving vowel identity [109].

A phonetic quality of vowel can be characterized by resonance peaks (formants) of the vocal-tract transfer function. It is well known that lower two or three formant frequencies provide effective and probably enough information to classify phonetic quality of vowel in acoustical analysis [103]. Thus, it is reasonable to suppose that the formant frequencies might be crucial cue for vowel perception. The psychoacoustical experiments using synthesized vowel stimuli

indicated that formant frequency changes were the most important dimensions that caused subjects to report a change in phonetic quality. Changes in spectral tilt and filtering pass-band/stopband, while clearly audible, did not induce the impression of a phonetic change – something else changed, such as the speaker or perceived transmission channel [110]. In actual, formant-based vowel perception model can predict listener's response to vowel-like signals in many cases.

Studies from [108] has shown that vowels having lower values of F1/F2 are identified as /u/ and with higher values as /e/. Some previous studies [111] has shown the importance of additional third formant (F3) serving as important cue for vowel perception. The same has been true in the perception of noise-excited vowels, where higher formants could play a key role [108]. Besides these, it was also concluded in [111] that the whole spectral shape encompassing these formants is also necessary for the perception of vowels. [112] studied vowel perception by applying Formant Centre of Gravity (FCOG) hypothesis to the back vowels.

According to this hypothesis, two formants can be merged and replaced by one single formant if the distance of separation between them is less than 3-3.5 Bark. Also, an increase in the relative amplitudes of both the formants may cause the centre of gravity to shift towards a higher frequency and vice-versa, thereby causing the spectral shape to differ accordingly. This spectral shape can be attributed to the rise or drop in F1/F2 frequencies, as they are related to the vowel height [113], [103], [114], [115], and vowel openness respectively. Apart from these formants, the fundamental frequency F0, also plays an important role in the perception of a vowel [116], [108], [117].

The pattern of resonances depends on F0 [118] and as F0 increases, the formant frequencies also change in order to maintain the vowel identity [117], [119], The vowels synthesized on same F0 are less intelligible compared to the vowels synthesized on different fundamental frequencies [120]. Also vowel identification can be accounted to the context in which they are perceived. Previous studies on vowel perception have revealed that isolated vowels are poorly identified when compared to the vowels present in the consonantal context [121], specifically when surrounded by stop consonants [122]. Isolated vowels had high perceptual identification error rates when compared to the vowels in CVC cluster [123]. [124] proved that vowels can be well-identified perceptually even in the absence of context.

F1 frequency in front vowels can be estimated as the weighted mean of the most prominent harmonics. However, such a simple strategy will not work in the case of back vowels, because F1 and F2 are often close together in frequency, making it impossible to assign harmonics uniquely to F1 or F2. Estimation of F1 and F2 frequencies from resolved harmonics in back vowels would require a complex partitioning of the energy between the formants. Some researchers have suggested that a different form of processing may be involved when two formants are close together in frequency. The formant centre of gravity (FCOG) hypothesis [106], [125], maintains that two closely spaced formants are effectively merged into a single spectral prominence whose

centre of gravity determines the phonetic quality of the vowel. According to this hypothesis, the centre of gravity depends not only on the frequencies of both formants, but also on their relative amplitudes.

A basic but yet unsolved problem in acoustic phonetics is to express vowel quality in simple terms, having a unique relation to vowel identity. Expressing vowel quality in relation to vowel identity is one of the important problems in acoustic phonetics. Normally vowel space is reduced to a plane showing the positions of first two formants, but this information is nnot sufficient to express vowel identity. The higher formants have considerable influence on the identity, especially for the close front vowels. [126].

Also, the direct proportionality between formant level and formant frequency has also been claimed to play an important role in the perception of vowels. Normally, the vowel space is reduced to a plane showing the position of the first two formants, but this information is not sufficient to express vowel identity. The higher formants have considerable influence on the identity, especially the close front vowels [126].

This chapter is organized as follows: Section 5.1 gives the description of Linear Prediction Analysis in detail. Also, the description of signals is covered in this section. The experimental details such as description of database, assessment criteria and listening tests are mentioned in Sec. 5.3. Section 5.4 discusses the results obtained. The summary and conclusions are given in Sec. 5.5.

5.1 Autocorrelation method of Linear Predictive Analysis

The overview of Linear Predictive Analysis was presented in Sec. 4.1.1. However, this section presents a detailed discussion on LP Analysis, LP Synthesis and autocorrelation method. This is explained here, as this chapter presents work, where the source and system components of the speech signal are separated and then speech is synthesized back through linear predictive analysis. As mentioned previously, LP is used to study the contribution of both source and system components. This method is based on the source-filter model of speech production. The transfer function of the vocal tract system is modeled as an all-pole response, resulting from measuring the correlation between the samples. The method of linear prediction attempts to model the system transfer function as an all-pole model based on a least squares fit. The analysis filter is obtained on the basis of least squares optimization. The error signal is obtained as a difference in actual and predicted signals. In this model, the present speech sample is predicted as a linear combination of past samples.

The predicted signal is defined by the following equation

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{5.1}$$

where 'p' is the LP order.

5.1.1 LP Analysis

LP Analysis filter takes this speech sample as an input and gives the error signal (also known as LP residual) along with the coefficients. The error signal is given by:

$$e(n) = s(n) - \hat{s}(n) \tag{5.2}$$

The original signal $s(n)$ can be reconstructed from the predicted signal by the following relation

$$\hat{s}(n) = e(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{5.3}$$

Taking Z-transform on both the sides of Eq. 5.3 and deriving expression for $E(z)$ would result in

$$E(z) = [1 - \sum_{i=1}^{p} a_k z^{-k}] S(z) \tag{5.4}$$

where

$$1 - \sum_{i=1}^{p} a_k z^{-k} = A(z) \tag{5.5}$$

$A(z)$ represents the filter function of the analysis filter. The minimizing procedure of the error leads to the auto-correlation function given by:

$$-R(i) = \sum_{k=1}^{p} a_k R(i-k) \tag{5.6}$$

substituting for $i = 1, k = 1, i = 2, k = 2$ and so on, we get:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots R(p-1) \\ R(1) & R(0) & R(1) & \dots R(p-2) \\ \vdots & \vdots & \vdots & \ddots \quad \vdots \\ R(p-1) R(p-2) R(p-3) \dots & & & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix}$$

$$R.A = -r \tag{5.7}$$

$$A = -R^{-1}.r \tag{5.8}$$

$$\text{where} \quad A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \tag{5.9}$$

where A is the set of LP coffiecients that can be used to obtain the spectrum. This spectrum would be modelled as an all-pole filter system that ensures stability of the filter.

5.1.2 LP Synthesis

The LP synthesis method uses the inverse filter and the error obtained using analysis step to synthesize speech signal. The inverse filter is obtained by inverting the filter obtained containing autocorrelation coefficients given by A in Eq. 5.9. The transfer function of this inverse filter is given by $H(z)$,

$$H(z) = \frac{1}{A(z)} \tag{5.10}$$

$$= \frac{1}{1 + [\sum_{i=1}^{p} a_k z^{-k}]} \tag{5.11}$$

The output of the synthesis filter would be the reconstructed signal $s(n)$.

5.1.3 Description of the signals

The LP analysis decomposes the speech signal $s(n)$ shown in Fig. 5.1(a) into source and system components. The source is the error component $e(n)$ given by Eq. 5.1.1 obtained as the difference in actual vs predicted speech, as shown in Fig. 5.1(b) obtained as a result of passing the speech signal through the analysis filter model. Formants representing the VTS are shown in Fig. 5.1(c), and are realized as an all-pole filter function denoted by Eq. (5.11), is estimated over a block of 20 ms, based on minimizing the mean square estimation. The Fig. 5.1(d) displays the waveform for a synthesized speech segment obtained using the all pole model excited by a random noise. The resynthesized signal using the all pole model and error signal obtained from a frame of original speech signal is shown in Fig. 5.1(e).

5.2 Proposed Method

The proposed method for this study can be explained with the help of the Fig. 5.2. LP Analysis-by-synthesis method discussed in Sections 5.1.1 and 5.1.2 was used to separate excitation source (ES) and vocal tract system (VTS) components of the speech signal. The autocorrelation method of LP explained in Sec. 5.1.1 helps in extraction of LP coefficients which are used to obtain the spectrum. The LP spectrum is modelled as an all-pole filter.

LP synthesis method incorporates the inverse filter to resynthesize the signal. Three different cases were carried out here. Firstly, as can be seen from Fig. 5.2(a) representing the original case, in which the VTS component of all the five vowels were excited with the ES component of the same five vowels, i.e VTS of /a/ excited with ES of /a/, VTS of /e/ with ES of /e/ and so on. Figure 5.2(b) shows mixed case. As the name itself implies, the VTS and ES components of vowels were mixed. For instance, the VTS of /a/ is excited with ES of the rest four vowels /e/ /i/, /o/ and /u/. Similarly, the same is carried for other four vowels. Speech segments having

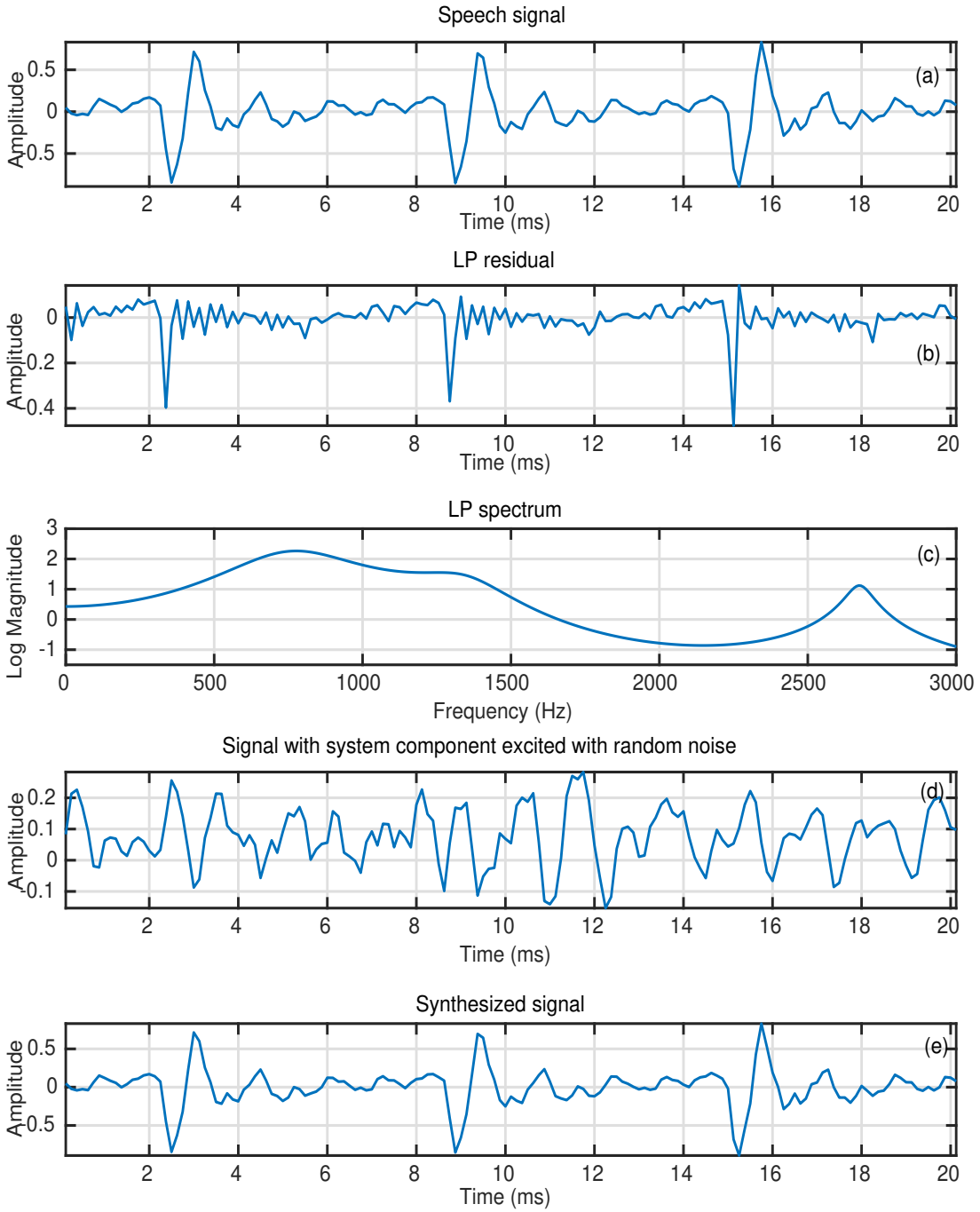

Figure 5.1: (a) Speech segment, (b) LP residual, (c) LP spectrum, (d) signal obtained from an all pole model excited with random noise and (e) signal obtained using filtering error signal through system response obtained from speech.

only the ES components of the vowels are shown in Fig. 5.2(c). Fig. 5.2(d) illustrates vowels with only the VTS component excited with random noise (RN).

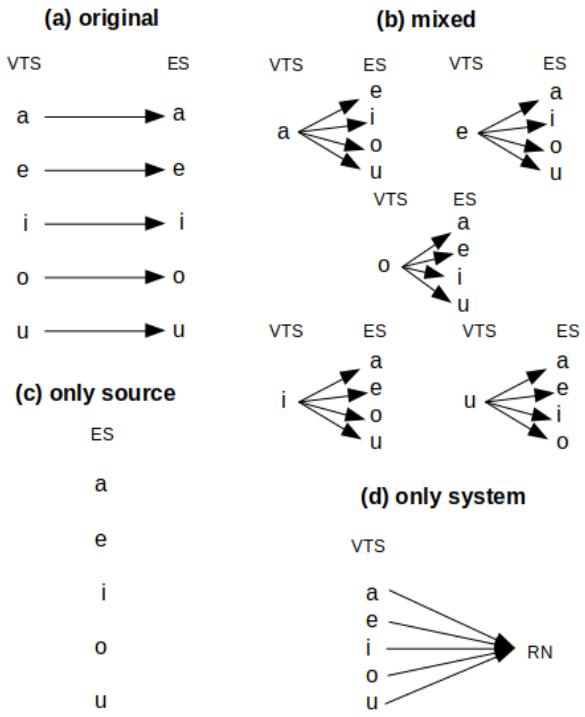

Figure 5.2: Illustration of proposed method

5.3 Experimental Details

5.3.1 Database

The size of the database collected was distributed across 50 speakers (25 male and 25 female), whose ages ranged from 18 to 38 years. The speakers were students of International Institute of Information Technology, Hyderabad (IIIT-H), India. The speakers were asked to utter 5 long vowel sounds (/a/, /e/, /i/, /o/, /u/) in isolation. These vowel sounds are taken from the International Phonetic Alphabet (IPA) vowel chart (shown in 1.8) denoted for the underlined parts in the words, a<u>r</u>m, f<u>a</u>ke, h<u>ea</u>t, r<u>oa</u>d and s<u>oo</u>t respectively. The recordings took place in the recording studio (silent recording room) of IIIT-H. The recordings were recorded using Audacity audio recorder and editor through Roland Octa-Capture USB Audio Interface connected to a microphone. The microphone used for the recording was set about a distance ranging from 10-15 cms from the speakers' mouth. The sampling rate of the recordings was maintained at 48 kHz with 16-bit quantization which were then downsampled to 16 kHz for the analysis.

5.3.2 Assesment criteria

The subjects chosen for the listening tests were a total of 16 male and 16 female speakers, who ranged from 20-60 years of age, out of which 15 were the speakers used for the database

collection. Some of the speakers were posing as listeners as this would produce few errors when their own articulations were played randomly during the listening tests [127].

5.3.3 Listening tests

Each of these 32 listeners was presented the speech files for 5 vowel sounds of 20 (10-male and 10-female) randomly selected speakers from the database. Therefore, each listener was presented with a total of 640 samples. The listeners carried on the perceptual tests using a pair of headphones. The listeners were made to listen to the vowel sounds and write down the symbol of the perceived vowel. In case, the listeners were unable to map any vowel symbol to that of the sound played, then they were asked to simply assign it a cross (x) mark, which means that no sound is perceived. This mark would make several assumptions such as, the sound was not audible enough (this may happen when the subject is more aged), the sound may not be clear, the sound may contain noise, or its an unrecognizable or confusing sound.

5.4 Results and Discussion

The results shown in Figure 5.3 are the confusion matrices plotted by evaluating the listening tests across 32 subjects. The confusion plots have the entries of the vowel sounds played (y-axis) versus the perceived (predicted) vowel sound (x-axis). Thus, the maximum entry would be 640. It was seen that the vowels approaching this number were perceived better.

Figure 5.3(a) shows the confusion matrix obtained for the samples synthesized using the source signal and system response obtained from the same vowel sounds. It simply means that the vowel sounds were synthesized using their own components. This was merely done to eliminate any bias, if existed, towards perception of a certain sound. It can be seen from Fig. 5.3(a) that the darker colour along the diagonal indicates that the vowels were correctly perceived as these were the results plotted for the synthesized files comprising both source and system components. Fig. 5.3(b) shows the confusion matrix for mixed segments synthesized using the system response of /a/ excited with the source components of /e/, /i/, /o/ and /u/. The column along vowel /a/ is more dominant indicating that the vowel /a/ is perceived as itself for majority of the listeners, signifying the importance of the system transfer function towards perception. Similar is the case for Figs. 5.3(c), 5.3(d), 5.3(e) and 5.3(f), where the vowels /e/, /i/, /o/ and /u/ are perceived mostly as themselves for most of the listeners, showing that the same vowel is perceived when its system component is interchanged with source of the rest. Also for the Fig. 5.3(g), the diagonal is dominant, but /i/ and /u/ are less perceived compared to others. Few listeners identified the vowel /i/ as /e/ and /u/ as /o/ here. Fig. 5.3(h) exhibits same behaviour as that of Fig. 5.3(a), with the system component contributing a major role in perception of these vowels.

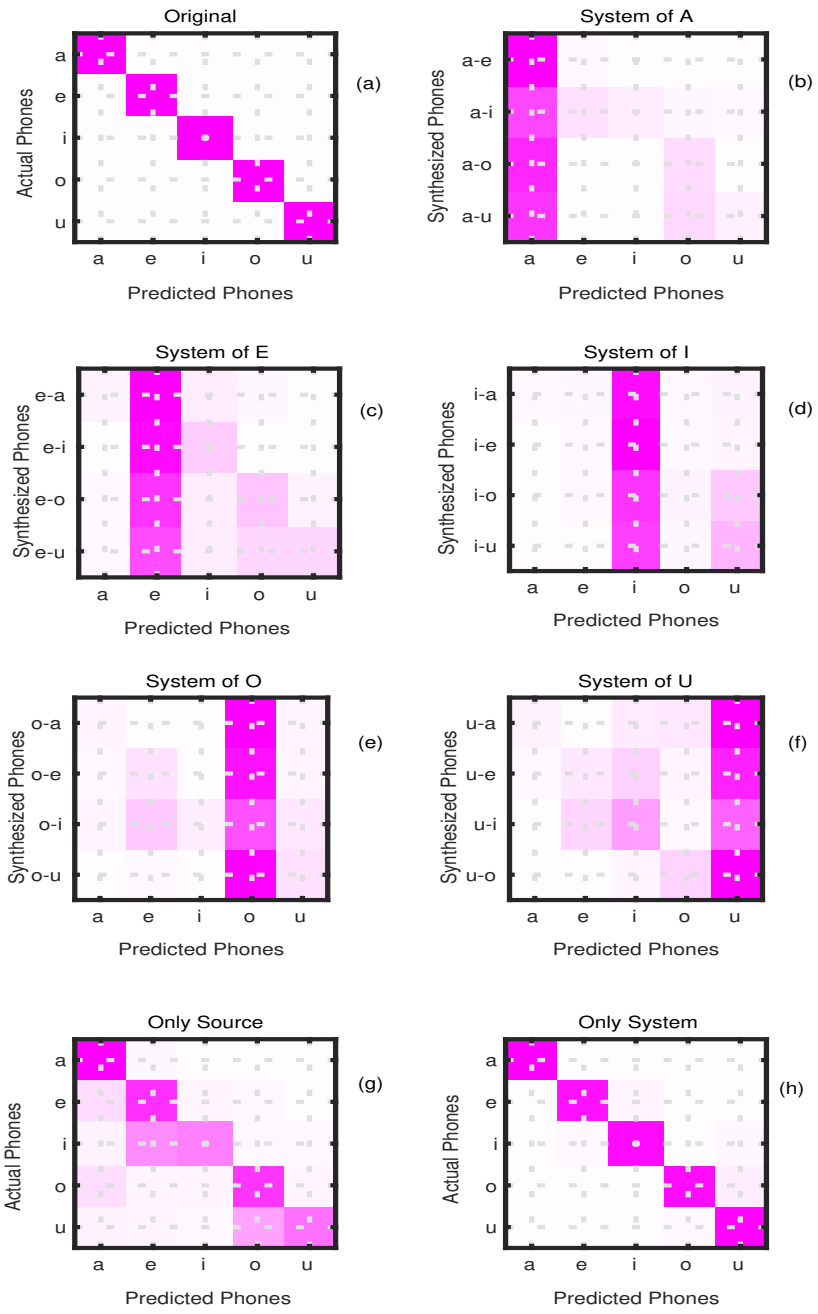

Figure 5.3: Confusion matrices for listening test results of (a) synthesized speech having source signal and system response obtained from the same vowel sounds, (b) mixed speech segments having system component of /a/ excited with source components of /e/,/i/,/o/,/u/, (c) mixed speech segments having system of /e/ excited with source of /a/,/i/,/o/,/u/, (d) mixed speech segments having system of /i/ excited with source of /a/,/e/,/o/,/u/, (e) mixed speech segments having system of /o/ excited with source of /a/,/e/,/i/,/u/, (f) mixed speech segments having system of /u/ excited with source of /a/,/e/,/i/,/o/, (g) speech segments with only the source components of the vowels /a/,/e/,/i/,/o/,/u/ and (h) speech segments having the system components of /a/,/e/,/i/,/o/,/u/ excited with random noise

It was observed that the transfer function of the vocal tract filter had a dominant role in the perception of vowels. Few listeners identified the vowel /i/ as /e/ and /u/ as /o/ here. This may be due to the fact that the vowels /i/ and /e/ are front vowels appearing on the same side of the vowel tract and back vowels /o/ and /u/ appearing on the same side as well. So the system transfer function does not vary much during production and as a result, they may sound similar. The results from Fig. 5.3(h) had comments from listeners saying that the synthesized files had a whispering effect (due to addition of whispering effect), but yet were clear enough for the perception level to keep high compared to those of results in Fig. 5.3(g) for the vowels synthesized with system component excited with random noise. The overall result would be attributed to the spectral characteristics of the vowels as they very well define the distribution of the formants, their slopes, thereby making the vocal tract system information significant in the perception of vowels.

5.5 Summary and conclusions

The role of source and system components was investigated towards the perception of vowels. Source and system information of vowel utterance is obtained using Linear Prediction (LP) analysis. Three listening experiments are conducted, where only the excitation source signals, VT response of a given vowel excited with random noise, excitation information of a certain vowel and VT information of a different vowel were used to synthesize speech signals. Results obtained from these experiments showed that vocal tract system information plays a significant and decisive role in the perception of vowels.

Chapter 6

Summary and conclusions

The study of this thesis mainly focussed on applying acoustic correlates in Speech Production, Perception and Speech Analysis. Acoustic correlates are the main features of speech sounds that distinguish sounds from one another. The work of thesis studies acoustic correlates of vowels and their relation to vowel categorization. Major part is explored for vowel sounds. Minor part of first study covers the use of one of the acoustic correlate for nasal sounds. Acoustic cues such as Formants, VSA, Formant Bandwidth were used in this thesis work.

The first study detects formant bandwidth for both oral and nasal tracts over the closed phase of the glottis. This is accomplished using vowels for oral tract and nasal sounds for nasal tract. ZTW method was computed to extract DRFs from HNGD spectra. The results obtained exhibit that nasals have higher bandwidth, due to higher impedance of the nasal tract when compared to vowels. This study can be further extended to detect regions of nasalization in vowels. Also it can be derived for nasalized vowels. This is because the nasalized vowels can incur the effects of nasalization which can possibly decrease their bandwidth further according to this study. It can also be incorporated to detect the actual bandwidth of oral and nasal cavitites by considering multiple resonance model.

The second study illustrates the comparison of speech performance among categories of gender viz., men, women and children (boys and girls) using VSA as a metric. Although VS has been used in speech pathology and speech disorders in past, but was never used to study the speech performance among genders. VS was constructed using the detected formants derived from LP method. So, this study again exploits two important acoustic correlates of vowels, which are formants and VSA. An ideal speech performance is considered to form quadrilateral (trapezoid) in F1-F2 plane. The results obtained from this study have shown the articulation of adults forming a qualdrilateral in F1-F2 plane. But this was not true in the case of children. The VS of children did not form a distinct quadrilateral due to improper vowel articulations resulting from lack of clear open-close vocal tract configurations. This study concluded that vowel [i] is articulated well by all speakers and vowels that have more dynamic articulations are [æ] and [ɒ]. This study validates the relationship between acoustic and articulatory characteristics of

speech influencing VS across gender and age. This work can be further extended to detect lip and tongue disorders, oral (mouth) cancer, autism and so on.

The objective of the third study was to analyze the role of excitation source and vocal tract system in the perception of vowels. This was investigated using LP analysis. Again, formants were used to represent the resonances of the vocal tract system, since, speech perception or the auditory system perceive sounds having different acoustic properties in distinctive ways. Three listening experiments are conducted, where only the excitation source signals, VT response of a given vowel excited with random noise, excitation information of a certain vowel and VT information of a different vowel were used to synthesize speech signals. Results obtained from these experiments showed that vocal tract system information plays a significant and decisive role in the perception of vowels. This work can be further extended in applying the same study to the cardinal vowel set in place of long vowels. This can show whether all the cardinal vowels (corner vowels) can be very well perceived with both the source and system components or either of them, thereby validating acoustic-auditory properties of vowels.

List of Publications

Publications related to thesis

1. **Haala Deeba Abbas**, RaviShankar Prasad, Bhanu Teja Nellore, and Suryakanth V. Gangashetty, "Study of Closed Phase Resonance Bandwidths for Oral and Nasal Tracts Using Zero Time Windowing", in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020, Barcelona, Spain, pp. 7369-7373.

Bibliography

[1] P. Ladefoged and K. Johnson. *A Course in Phonetics*. Cengage Learning, 2010.

[2] P. Ladefoged. *Vowels and Consonants*. Wiley, 2005.

[3] Kiran Lakkaraju, Samarth Swarup, and Les Gasser. Consensus under constraints: Modeling the great english vowel shift. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 1–8. Springer, 2012.

[4] James M Pickett. *The sounds of speech communication: A primer of acoustic phonetics and speech perception*. Univ Park Press, 1980.

[5] James Robert Glass. Nasal consonants and nasalized vowels: An acoustic study and recognition experiment. Master's thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering, 1984.

[6] K Sri Rama Murty, Bayya Yegnanarayana, and M Anand Joseph. Characterization of glottal activity from speech signals. *IEEE signal processing letters*, 16(6):469–472, 2009.

[7] Lawrence R Rabiner. *Digital processing of speech signals*. Pearson Education India, 1978.

[8] Ettien Koffi. The acoustic correlates of [±atr] vowels: An analysis by reference levels of anyi vowels. *Linguistic Portfolios*, 5(1):9, 2016.

[9] Ratree Wayland and Allard Jongman. Acoustic correlates of breathy and clear vowels: The case of khmer. *Journal of Phonetics*, 31(2):181–201, 2003.

[10] Stephen A Zahorian and Amir Jalali Jagharghi. Spectral-shape features versus formants as acoustic correlates for vowels. *The Journal of the Acoustical Society of America*, 94(4):1966–1982, 1993.

[11] Masashi Ito, Keiji Ohara, Akinori Ito, and Masafumi Yano. Relative importance of formant and whole-spectral cues for vowel perception. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[12] Peter Denes. Effect of duration on the perception of voicing. *The Journal of the Acoustical Society of America*, 27(4):761–764, 1955.

[13] Gordon E Peterson. The information-bearing elements of speech. *The Journal of the Acoustical Society of America*, 24(6):629–637, 1952.

[14] Ian H Witten. Driving the votrax speech synthesizer from a wide phonetic transcription with high-level prosodic markers. *International Journal of Man-Machine Studies*, 16(4):393–403, 1982.

[15] James Hillenbrand, Laura A Getty, Michael J Clark, and Kimberlee Wheeler. Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, 97(5):3099–3111, 1995.

[16] Shimon Sapir, Lorraine O Ramig, Jennifer L Spielman, and Cynthia Fox. Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech. *Journal of Speech, Language, and Hearing Research*, 53(1):114–125, 2010.

[17] Vaishna Narang and Deepshikha Misra. Acoustic space, duration and formant patterns in vowels of bangkok thai. *Int. J. of Asian Lang. Proc.*, 20(3):123–140, 2010.

[18] Kris Tjaden, Deanna Rivera, Gregory Wilding, and Greg S Turner. Characteristics of the lax vowel space in dysarthria. *Journal of Speech, Language, and Hearing Research*, 48(3):554–566, 2005.

[19] Gunnar Fant. *Speech sounds and features.* The MIT Press, 1973.

[20] Ray D Kent and Y-J Kim. Toward an acoustic typology of motor speech disorders. *Clinical linguistics & phonetics*, 17(6):427–445, 2003.

[21] Peter Ladefoged. *Elements of acoustic phonetics.* University of Chicago Press, 1996.

[22] Lindblom Bjorn. *Explaining phonetic variation: A sketch of the H&H theory.* Springer, 1990.

[23] Sarah Hargus Ferguson and Hugo Quené. Acoustic correlates of vowel intelligibility in clear and conversational speech for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 135(6):3570–3584, 2014.

[24] P. Skandera and P. Burleigh. *A Manual of English Phonetics and Phonology: Twelve Lessons with an Integrated Course in Phonetic Transcription.* Narr, 2005.

[25] M. Celce-Murcia, D. M. Brinton, and J. M. Goodwin. *Teaching Pronunciation: A Reference for Teachers of English to Speakers of Other Languages.* Cambridge University Press, 1996.

[26] RaviShankar Prasad, Sudarsana Reddy Kadiri, Suryakanth V Gangashetty, and Bayya Yegnanarayana. Discriminating nasals and approximants in english language using zero time windowing. In *Proceedings of the International Conference on Spoken Language Processing*, pages 177–181, 2018.

[27] Hisayoshi Suzuki, Takayoshi Nakai, Jianwu Dang, and Chengxiang Lu. Speech production model involving subglottal structure and oral-nasal coupling through closed velum. In *ICSLP*, volume 90, pages 437–440, 1990.

[28] Kenneth N Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.

[29] Osamu Fujimura. Analysis of nasal consonants. *The Journal of the Acoustical Society of America*, 34(12):1865–1875, 1962.

[30] Marilyn Y Chen. Nasal detection module for a knowledge-based speech recognition system. In *Sixth International Conference on Spoken Language Processing*, 2000.

[31] J. Glass and V. Zue. Detection of nasalized vowels in american english. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, volume 10, pages 1569–1572. IEEE, 1985.

[32] Paul Mermelstein. On detecting nasals in continuous speech. *The Journal of the Acoustical Society of America*, 61(2):581–587, 1977.

[33] Tarun Pruthi and Carol Y Espy-Wilson. Acoustic parameters for automatic detection of nasal manner. *Speech Communication*, 43(3):225–239, 2004.

[34] Clifford Weinstein, Stephanie S McCandless, Lee Mondshein, and Victor Zue. A system for acoustic-phonetic analysis of continuous speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):54–67, 1975.

[35] Gunnar Bjuggren and Gunnar Fant. The nasal cavity structures. *STL-QPSR*, 5(4):5–7, 1964.

[36] Joseph M Anand, S Guruprasad, and B Yegnanarayana. Extracting formants from short segments of speech using group delay functions. In *Ninth International Conference on Spoken Language Processing*, 2006.

[37] Leon Cohen, Khaled Assaleh, and Adam Fineberg. Instantaneous bandwidth and formant bandwidth. In *Statistical Signal and Array Processing, 1992. Conference Proceedings., IEEE Sixth SP Workshop on*, pages 13–17. IEEE, 1992.

[38] Alexandros Potamianos and Petros Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *Journal of Acoustical Society of America*, 99(6):3795–3806, 1996.

[39] Yanli Zheng and Mark Hasegawa-Johnson. Particle filtering approach to bayesian formant tracking. In *Statistical Signal Processing, IEEE Workshop on*, pages 601–604. IEEE, 2003.

[40] N Reddy and M Swamy. High resolution formant extraction from linear-prediction phase spectra. *IEEE Trans. Audio, Speech Lang. Process.*, 32(6):1136–1144, 1984.

[41] O Yasojima, Y Takahashi, and M Tohyama. Resonant bandwidth estimation of vowels using clustered-line spectrum modeling for pressure speech waveforms. In *Signal Processing and Information Technology, IEEE International Symposium on*, pages 589–593. IEEE, 2006.

[42] Yegnanarayana Bayya and Dhananjaya N Gowda. Spectro-temporal analysis of speech signals using zero time windowing and group delay function. *Speech Communication*, 55(6):782–795, 2013.

[43] Alan V Oppenheim and Ronald W Schafer. Digital signal processing(book). *Research supported by the Massachusetts Institute of Technology, Bell Telephone Laboratories, and Guggenheim Foundation. Englewood Cliffs, N. J., Prentice-Hall, Inc., 1975. 598 p*, 1975.

[44] Jilt Sebastian, Manoj Kumar, and Hema A Murthy. An analysis of the high resolution property of group delay function with applications to audio signal processing. *Speech Communication*, 81:42–53, 2016.

[45] Anand Joseph Xavier Medabalimi, Guruprasad Seshadri, and B Yegnanarayana. Extraction of formant bandwidths using properties of group delay functions. *Speech Communication*, 63:70–83, 2014.

[46] RaviShankar Prasad and Bayya Yegnanarayana. Robust pitch estimation in noisy speech using ztw and group delay function. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[47] Ravi Shankar Prasad and B Yegnanarayana. Determination of glottal open regions by exploiting changes in the vocal tract system characteristics. *Journal of Acoustical Society of America*, 140(1):666–677, 2016.

[48] RaviShankar Prasad and Bayya Yegnanarayana. Identification and classification of fricatives in speech using zero time windowing method. In *INTERSPEECH*, pages 187–191, 2018.

[49] RaviShankar Prasad. Analysis of dynamics of vocal tract system using zero time windowing method. Technical report, International Institute of Information Technology Hyderabad, 2019.

[50] RaviShankar Prasad and B Yegnanarayana. A study of vowel nasalization using instantaneous spectra. *Computer Speech & Language*, 69:101214, 2021.

[51] Paavo Alku, Juha Vintturi, and Erkki Vilkman. On the linearity of the relationship between the sound pressure level and the negative peak amplitude of the differentiated glottal flow in vowel production. *Speech communication*, 28(4):269–281, 1999.

[52] Roel Smits and B Yegnanarayana. Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing*, 3(5):325–333, 1995.

[53] Yannis Stylianou. Removing linear phase mismatches in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(3):232–239, 2001.

[54] Bayya Yegnanarayana, K Sri Rama Murty, and S Rajendran. Analysis of stop consonants in indian languages using excitation source information in speech signal. In *Proc. Workshop Speech Anal. Process. Knowledge Discovery*, pages 4–6, 2008.

[55] K Sri Rama Murty and B Yegnanarayana. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1602–1613, 2008.

[56] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[57] Bayya Yegnanarayana. Formant extraction from linear-prediction phase spectra. *Journal of Acoustical Society of America*, 63(5):1638–1640, 1978.

[58] Bayya Yegnanarayana and Hema A Murthy. Significance of group delay functions in spectrum estimation. *IEEE Transactions on signal processing*, 40(9):2281–2289, 1992.

[59] RaviShankar Prasad and B Yegnanarayana. Acoustic segmentation of speech using zero time liftering (ztl). In *INTERSPEECH*, pages 2292–2296, 2013.

[60] Huei-Mei Liu, Feng-Ming Tsao, and Patricia K Kuhl. The effect of reduced vowel working space on speech intelligibility in mandarin-speaking young adults with cerebral palsy. *The Journal of the Acoustical Society of America*, 117(6):3879–3889, 2005.

[61] Benjamin Picart, Thomas Drugman, and Thierry Dutoit. Analysis and synthesis of hypo-and hyperarticulated speech. In *Proc. Seventh ISCA Workshop on Speech Synthesis (SSW)*, 2010.

[62] Sri Harsha Dumpala, Bhanu Teja Nellore, Raghu Ram Nevali, Suryakanth V. Gangashetty, and B. Yegnanarayana. Robust vowel landmark detection using epoch-based features. In *Interspeech 2016*, pages 160–164, 2016.

[63] Hamidreza Baradaran Kashani, Abolghasem Sayadiyan, and Hamid Sheikhzadeh. Vowel detection using a perceptually-enhanced spectrum matching conditioned to phonetic context and speaker identity. *Speech Communication*, 91:28–48, 2017.

[64] Brad H Story and Kate Bunton. Vowel space density as an indicator of speech performance. *The Journal of the Acoustical Society of America*, 141(5):458–464, 2017.

[65] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468, 1999.

[66] Ann R Bradlow, Gina M Torretta, and David B Pisoni. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20(3-4):255–272, 1996.

[67] Kate Bunton and Mark Leddy. An evaluation of articulatory working space area in vowel production of adults with down syndrome. *Clinical linguistics & phonetics*, 25(4):321–334, 2011.

[68] Amy T Neel. Vowel space characteristics and vowel identification accuracy. *Journal of Speech, Language, and Hearing Research*, 51(3):574–585, 2008.

[69] Chris Davis and Jeesun Kim. Is speech produced in noise more distinct and/or consistent? *Speech Science and Technology*, pages 46–49, 2012.

[70] Sarah Hargus Ferguson. *Vowels in clear and conversational speech: Talker differences in acoustic features and intelligibility for normal-hearing listeners*. Indiana University, 2002.

[71] Valerie Hazan and Rachel Baker. Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style? In *DiSS-LPSS Joint Workshop 2010*, 2010.

[72] Alexander Kain, Akiko Amano-Kusumoto, and John-Paul Hosom. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *The Journal of the Acoustical Society of America*, 124(4):2308–2319, 2008.

[73] CM Higgins and MM Hodge. Vowel area and intelligibility in children with and without dysarthria. *Journal of Medical Speech-Language Pathology*, 10(4):271–277, 2002.

[74] Robert A Prosek, Allen A Montgomery, Brian E Walden, and David B Hawkins. Formant frequencies of stuttered and fluent vowels. *Journal of Speech, Language, and Hearing Research*, 30(3):301–305, 1987.

[75] Gary Weismer, Jacqueline S Laures, Jing-Yi Jeng, Ray D Kent, and Jane F Kent. Effect of speaking rate manipulations on acoustic and perceptual aspects of the dysarthria in amyotrophic lateral sclerosis. *Folia Phoniatrica et Logopaedica*, 52(5):201–219, 2000.

[76] Tara L Whitehill, Valter Ciocca, Judy C-T Chan, and Nabil Samman. Acoustic analysis of vowels following glossectomy. *Clinical linguistics & phonetics*, 20(2-3):135–140, 2006.

[77] Wolfram Ziegler and D Von Cramon. Spastic dysarthria after acquired brain injury: An acoustic study. *British Journal of Disorders of Communication*, 21(2):173–187, 1986.

[78] W Ziegler and D Von Cramon. Vowel distortion in traumatic dysarthria: A formant study. *Phonetica*, 40(1):63–78, 1983.

[79] Paul A McRae, Kris Tjaden, and Barbra Schoonings. Acoustic and perceptual consequences of articulatory rate change in parkinson disease. *Journal of Speech, Language, and Hearing Research*, 45(1):35–50, 2002.

[80] Stefan Scherer, Louis-Philippe Morency, Jonathan Gratch, and John Pestian. Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4789–4793, 2015.

[81] Stefan Scherer, Gale M Lucas, Jonathan Gratch, Albert Skip Rizzo, and Louis-Philippe Morency. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, 7(1):59–73, 2016.

[82] Kris Tjaden and Gregory E Wilding. Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*, 47(4):766–783, 2004.

[83] Young Kang, Kyu-Chul Yoon, Hak-Seung Lee, and Cheol-Jae Seong. A comparison of parameters of acoustic vowel space in patients with parkinson's disease. *Phonetics and Speech Sciences*, 2(4):185–192, 2010.

[84] Harlan Lane, Melanie Matthies, Joseph Perkell, Jennell Vick, and Majid Zandipour. The effects of changes in hearing status in cochlear implant users on the acoustic vowel space and cv coarticulation. *Journal of Speech, Language, and Hearing Research*, 44(3):552–563, 2001.

[85] Ann R Bradlow. A comparative acoustic study of english and spanish vowels. *The Journal of the Acoustical Society of America*, 97(3):1916–1924, 1995.

[86] Shimon Spair, Jennifer Spielman, Lorraine O Ramig, Stephanie L Hinds, Stefanie Countryman, Cynthia Fox, and Brad Story. Effects of intensive voice treatment (the lee silverman voice treatment [lsvt]) on ataxic dysarthria: A case study. *American Journal of Speech-Language Pathology*, 12(4):387–399, 2003.

[87] Sapir Shimon, Spielman Jennifer L, Ramig Lorraine O, Story Brad H, and Fox Cynthia. Effects of intensive voice treatment (the lee silverman voice treatment [lsvt]) on vowel articulation in dysarthric individuals with idiopathic parkinson disease: acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*, 50(4):899–912, 2007.

[88] Moura Carla Pinto, Cunha Luis Miguel, Vilarinho Helena, Cunha Maria Joao, Freitas Diamantino, Palha Miguel, Pueschel M, and Pais-Clemente M. Voice parameters in children with down syndrome. *Journal of Voice*, 22(1):34–42, 2008.

[89] Michiko Hashi, John R Westbury, and Kiyoshi Honda. Vowel posture normalization. *The Journal of the Acoustical Society of America*, 104(4):2426–2437, 1998.

[90] Byunggon Yang. A comparative study of american english and korean vowels produced by male and female speakers. *Journal of phonetics*, 24(2):245–261, 1996.

[91] Rosen Kristin M, Goozee Justine V, and Murdoch Bruce E. Examining the effects of multiple sclerosis on speech production: Does phonetic structure matter? *Journal of Communication Disorders*, 41(1):49–69, 2008.

[92] Yana Yunusova, Gary Weismer, John R Westbury, and Mary J Lindstrom. Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research*, 51(3):596–611, 2008.

[93] Ann R Bradlow and Tessa Bent. The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, 112(1):272–284, 2002.

[94] Ying-Chiao Tsao, Gary Weismer, and Kamran Iqbal. The effect of intertalker speech rate variation on acoustic vowel space. *The Journal of the Acoustical Society of America*, 119(2):1074–1082, 2006.

[95] Bishnu S Atal and Suzanne L Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The journal of the acoustical society of America*, 50(2B):637–655, 1971.

[96] Bishnu S Atal and MR Schroeder. Linear prediction analysis of speech based on a pole-zero representation. *The Journal of the Acoustical Society of America*, 64(5):1310–1318, 1978.

[97] DH Klatt. Acoustic theory of terminal analog speech synthesis. In *Proc. 1972 Int. Conf. Speech Communication and Processing*, pages 131–135, 1972.

[98] Ronald W Schafer and Lawrence R Rabiner. System for automatic formant analysis of voiced speech. *The Journal of the Acoustical Society of America*, 47(2B):634–648, 1970.

[99] John Makhoul. Spectral analysis of speech by linear prediction. *IEEE Transactions on Audio and Electroacoustics*, 21(3):140–148, 1973.

[100] P MERMELSTEIN. Speech synthesis with the aid of a recursive filter approximating the transfer function of the nasalized vocal tract(speech synthesis aided by recursive filter approximating transfer function of nasalized vocal tract). In *IEEE The 1972 Conf. on Speech Commun. and Process. p 152-156(SEE N 73-23119 14-07)*, 1972.

[101] Ursula Gisela Goldstein. *An articulatory model for the vocal tracts of growing children*. PhD thesis, Massachusetts Institute of Technology, 1980.

[102] Jason A Whitfield and Alexander M Goberman. Articulatory-acoustic vowel space: Associations between acoustic and perceptual measures of clear speech. *International Journal of Speech-Language Pathology*, 19(2):184–194, 2017.

[103] Gordon E Peterson and Harold L Barney. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184, 1952.

[104] Winifred Strange. Evolving theories of vowel perception. *The Journal of the Acoustical Society of America*, 85(5):2081–2087, 1989.

[105] Gordon E Peterson. Parameters of vowel quality. *Journal of Speech and Hearing Research*, 4(1):10–29, 1961.

[106] Pierre Delattre, Alvin M. Liberman, Franklin S. Cooper, and Louis J. Gerstman. An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns. *Word*, 8(3):195–210, 1952.

[107] Diana Deutsch. *Psychology of music*. Elsevier, 2013.

[108] Hiroya Fujisaki and Takako Kawashima. The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on audio and electroacoustics*, 16(1):73–77, 1968.

[109] Dennis Klatt. Prediction of perceived phonetic distance from critical-band spectra: A first step. In *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 1278–1281. IEEE, 1982.

[110] DE Klatt. Speech processing strategies based on auditory models. *The representation of speech in the peripheral auditory system*, 1982.

[111] Masashi Ito, Jun Tsuchida, and Masafumi Yano. On the effectiveness of whole spectral shape for vowel perception. *The Journal of the Acoustical Society of America*, 110(2):1141–1149, 2001.

[112] Peter F Assmann. The perception of back vowels: Centre of gravity hypothesis. *The Quarterly Journal of Experimental Psychology*, 43(3):423–448, 1991.

[113] Tsutomu Chiba and Masato Kajiyama. *The vowel: Its nature and structure*. Tokyo-Kaiseikan, 1941.

[114] Kopp Potter. Green: Visible speech. *New York*, 1947.

[115] Ralph K Potter and John C Steinberg. Toward the specification of speech. *The Journal of the Acoustical Society of America*, 22(6):807–820, 1950.

[116] W Ainsworth. Intrinsic and extrinsic factors in vowel judgments. *Auditory analysis and perception of speech*, pages 103–113, 1975.

[117] Roger L Miller. Auditory tests with synthetic vowels. *The Journal of the Acoustical Society of America*, 25(1):114–121, 1953.

[118] Dieter Maurer and Theodor Landis. Fo-dependence, number alteration, and non-systematic behaviour of the formants in German vowels. *International Journal of Neuroscience*, 83(1-2):25–44, 1995.

[119] Hartmut Traunmuller. The role of the fundamental and the higher formants in the perception of speaker size, vocal effort, and vowel openness. In *Proceedings of Franco-Swedish Seminar on Speech, Grenoble, France1985*, 1985.

[120] JPL Brokx and SG Nooteboom. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10(1):23–36, 1982.

[121] Terry L. Gottfried and Winifred Strange. Identification of coarticulated vowels. *The Journal of the Acoustical Society of America*, 68(6):1626–1635, 1980.

[122] Winifred Strange and Terry L Gottfried. Task variables in the study of vowel perception. *The Journal of the Acoustical Society of America*, 68(6):1622–1625, 1980.

[123] Ilse Lehiste and David Meltzer. Vowel and speaker identification in natural and synthetic speech. *Language and Speech*, 16(4):356–364, 1973.

[124] Peter F Assmann, Terrance M Nearey, and John T Hogan. Vowel identification: Orthographic, perceptual, and acoustic aspects. *The Journal of the Acoustical Society of America*, 71(4):975–989, 1982.

[125] Ludmilla A. Chistovich and Valentina V. Lublinskaya. The 'center of gravity'effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing research*, 1(3):185–195, 1979.

[126] Rolf Carlson, Bjorn Granstrom, and Gunnar Fant. Some studies concerning perception of isolated vowels. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 11(2-3):19–35, 1970.

[127] Arlene Earley Carney, Thomas Edman, Winifred Strange, and James J Jenkins. Advantage of speaker as listener in a vowel identification task. *The Journal of the Acoustical Society of America*, 73(6):2222–2223, 1983.