Novel Acoustic Features for Detection of Hypernasality in Cleft Palate Speech

Akhilesh Kumar Dubey

Novel Acoustic Features for Detection of Hypernasality in Cleft Palate Speech

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

Akhilesh Kumar Dubey

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

AUGUST 2020

Certificate

This is to certify that the thesis entitled "**Novel Acoustic Features for Detection of Hypernasality in Cleft Palate Speech**", submitted by **AKHILESH KUMER DUBEY** (146102013), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under our supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in our opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:
Guwahati.

Prof. S. R. Mahadeva Prasanna
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

Dated:
Guwahati.

Prof. S. Dandapat
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To

My parents

for their love, support and blessing

My wife and kids

for their support and sacrifice

Acknowledgements

I express my deep and sincere gratitude to my research supervisors, Prof. S. R. M. Prasanna and Prof. S. Dandapat for providing me an opportunity to work under their guidance. I am thankful for their continuous scholarly guidance in all aspects, motivation, and support throughout the doctoral studies. Their dedication, discipline, and hard work are the source of motivation for me. Without their support, it would be completely impossible for me to carry out the research work and bring the thesis to this level. I would also like to sincerely thank them for providing me with financial support for attending conferences and workshops. I am grateful to Prof. Rohit Sinha, the Chairman of the Doctoral Committee and Head, Department of Electronics and Electrical Engineering, IIT Guwahati for providing valuable suggestions on my work throughout the years. I am also thankful to my doctoral committee members Dr. Prithwijit Guha and Dr.Priyankoo Sarmah for their encouragement and valuable suggestions on my work. I am very much grateful to them for their insightful comments and constructive criticisms, which helped me bring my work to the current form. I would like to thank all the faculty members and office staff of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their help in carrying out this research work. I would also like to convey my gratitude to all technical and non-technical staff of the EEE department for their help and support throughout my Ph.D. Also, a heartful thanks to Dr. Abhishek Shrivastava of the Design Department for his valuable suggestions and encouragement for my research work.

This thesis would not become possible without the help and support of Speech-language pathologists (SLPs) and staff of All India Institute of Speech and Hearing (AIISH), Mysore. Especially, I would like to express my sincere thanks to Prof. M. Pushpavathi and Prof. Ajish K. Abraham for providing a wealth of knowledge about speech-language pathology. Their timely help and valuable suggestions helped me to formulate the objective of this thesis. I would like to acknowledge the help of Dr. Gopi Sankar, Dr. Navya, Mr. Gopi Kishor, Mrs. Deepthi, Ms. Nikitha, and Mr. Girish, during data collection and perceptual evaluation. Further, my thanks go to children, parents, and teachers for their cooperation during data collection. I owe my special thanks to Dr. Gayadhar Pradhan, Associate Professor, Electronics and Communication Department, NIT Patna, who created my interest in the speech signal processing subject during my M.Tech course and motivated me to pursue Ph.D.

My sincere thanks go to my seniors Dr. Deepak K.T., Mr. Ramesh, Dr. Nagraj Adiga, Dr.

Biswajit Dev Sarma, Dr. Banriskhem K. Khonglah, Dr. Rohan Kumar Das, Dr.Rajib Sharma, Dr. Bidisha Sharma, and Dr. Subhasis Mandal, for mentoring me during my research life. My special thanks to my closed friends Dr. Vikram CM, Dr. Sishir Kalita, and Protima Namo Sudro, for their support since from the beginning of my Ph.D. to thesis correction. Useful technical discussions with them shaped my research in many aspects. A thankful note for my lab-mates Moa, Himakshi, Sarfaraz, Shikha, Mrinmoy, Sandeep, Saswati, Balaji, Sriram, Sukanya, Vineeta, Prabhakar, Shoubhik, Deepika, Anik, Alex, Ato, Tilendra for their direct/indirect contributions during my stay at IIT Guwahati. I am thankful to my Ph.D. batch-mates Dr. Vivek, Mathew, Mohit, Ramanand Sagar, Vimal and Kaushik for helping me during the course work of my Ph.D. I am also thankful to Mrs. Sujata Sinha and Dr. Luke Horo for helping me during my paper correction. I convey my sincere thanks to the Ministry of Human Resources Department (MHRD), Govt. of India for providing fellowship during for Ph.D. Also, I would like to acknowledge the Department of Biotechnology (DBT), Govt. of India for providing financial assistance for the data collection.

I would like to thank Dr. Srikant Prasad and his family and Dr. Kelothu Suresh and his family for helping and supporting me and my family in the MSH hostel. I am very much thankful to my wife Mrs. Reena Dubey for encouraging me and handling all the responsibilities of family and children so that I can focus on my Ph.D. I thanks to my parents, sister, brother-in-law, brother, sister-in-law and rest of my family members for their blessings, support, encouragement and prayers for my success. Finally, I am thankful to goddess MAA KAMAKHYA for her blessing, because of which I could come to the end of my Ph.D. journey.

Akhilesh Kumar Dubey

Abstract

This thesis aims towards the development of an objective method for the assessment of hypernasality in cleft palate (CP) speech. The method is developed based on the hypernasality severity detection of /a/, /i/ and /u/ vowels present in CP speech. These vowels get nasalized due to the addition of nasal pole-zero pairs, as coupling of the nasal tract with the oral tract happens during the production of voiced sound in CP speech. The spectral analysis of nasalized vowels shows the presence of nasal formant and antiformant pairs in the vowel spectrum, especially in the vicinity of first formant (F_1), broadening of (F_1), and overall flattening of the spectrum. Hence, the spectral characteristics of hypernasal vowel deviate from the normal vowel in the form of centralization of energy in lower frequencies and modification in harmonics strength and envelope of the spectrum. In this thesis, firstly, temporal, sinusoidal model-based and cepstral features are explored for the normal vs. hypernasal vowel detection. Later, features are used for the hypernasality severity detection to develop an objective method for hypernasality assessment.

In the first work, hypernasality detection is attempted using the temporal features (time-domain processing based features): vocal tract constriction (VTC) and peak to side-lobe ratio (PSR). The VTC feature captures the low-frequency prominence in the signal and the PSR feature captures the residual signal characteristics around epoch locations. Hypernasality detection performance using the combined (VTC+PSR) feature is found better compared to the baseline features for all vowels. However, the performance is poor compared to the performance obtained using the Mel-frequency cepstral coefficients MFCC feature. In the second work, detection of hypernasality is performed using the sinusoidal model-based normalized harmonics amplitude (NHA), harmonics amplitude ratio (HAR) and prominent harmonics frequency (PHF) features. These features are based on the strength of harmonics. The analysis shows that the nature of these features is different for hypernasal vowels compared to normal vowels. The discriminative capability of each dimension of the features is measured by computing a statistical dependency (SD) measure between feature dimensions and class labels. The hypernasality detection performance using the combined (NHA+HAR+PHF) feature is better compared to the (VTC+PSR) feature for /a/ and /i/ vowels and is comparable for /u/ vowel. However, the performance is poor compared to the MFCC feature. In the third work, three cepstral features namely, Hilbert envelope of numerator of group delay function (HNGDF), pitch-adaptive Mel-frequency cepstral coefficients (PAMFCC) and spectral moment features augmented with low-order

cepstral coefficients (SMAC) features are used for the hypernasality detection. The HNGDF feature is the cepstral coefficients extracted from a high spectro-temporal HNGD spectrum. The PAMFCC feature is the cepstral coefficients extracted from the cepstral smooth spectrum. The SMAC feature is the concatenation of spectral moments of signals obtained after band pass filtering of speech signal with the lower order cepstral coefficients. Compared to the MFCC feature, the HNGDF feature gives poor performance for hypernasality detection, whereas the PAMFCC and SMAC features give better performance for all vowels. In the last, hypernasality detection is also performed using the inter combination of features. The combined feature (VTC+PSR+PAMFCC) performs best for /a/ and /i/ vowels, whereas combined feature (VTC+PSR+SMAC) performs best for /u/ vowel.

In the last work, a method for hypernasality severity detection based on the nasality score between [0 to 1] corresponding to the speaker's speech is proposed. The validation of the proposed method is also performed in this work by computing the nasality score corresponding to children's speech with normal, mild and moderate-severe hypernasality. The proposed method for hypernasality severity detection could be used for hypernasality assessment of CP children's speech.

Keywords: Hypernasality, cleft palate, vocal tract constriction, peak to side-lobe ratio, zero time windowing, first spectral moments, sinusoidal model, pitch adaptive liftering.

Contents

List of Figures

List of Tables

List of Acronyms

AIISH	All India Institute of Speech and Hearing
ANOVA	One-way analysis of variance
AR	Autoregression
ARMA	Autoregression moving average
AUROC	Area Under Receiver Operating Characteristic
BP	Backward propagation
CL	Cleft lip
CNN	Convolutional neural network
CP	Cleft Palate
CLP	Cleft Lip and Palate
CV	Consonant-vowel
CVCV	Consonant-vowel-consonant-vowel
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DHF	Dominant harmonics frequency
DNN	Deep neural network
EMD	Empirical Mode Decomposition
F1, F2, F3	First, second and third formant
GCIs	Glottal closure instants
GDAM	Group delay function based acoustic measure
GNE	Glottal to Noise Excitation Ratio
GUI	Graphical user interface
HAR	Harmonics amplitude ratio
HE	Hilbert envelope

HNGDF	Hilbert envelope of numerator of group delay spectrum function
HNR	Harmonic to Noise Ratio
LP	Linear prediction
LPCC	Linear prediction cepstral coefficient
MFCC	Mel-frequency cepstral coefficient
MFD	Modified group delay
NGD	Numerator of the group delay
NHA	Normalized harmonics amplitude
NLD	Nonlinear dynamic features
NNE	Normalized Noise Energy
PAMFCC	Pitch-adaptive Mel-frequency cepstral coefficient
PSR	Peak to side-lobe ratio
RBF	Radial basis function
ROC	Receiver Operating Characteristic
SD	Statistical dependency
SLPs	Speech-language pathologists
SMAC	Spectral moment features augmented by low-order cepstral coefficients
SVM	Support vector machine
SWLP	Stabilized weighted linear prediction
TEO	Teager energy operator
TEOF	Teager energy operator based feature
TONAR	The Oral Nasal Acoustic Ratio
VPD	Velopharyngeal Dysfunction
VSA	Vowel space area
VLHR	Voice low tone to high tone ratio
VTC	Vocal tract constriction
XLP	Extended weighted linear prediction
ZFFS	Zero frequency filtered signal
ZFR	Zero-frequency resonators
ZTW	Zero time windowing

1

Introduction

Contents

Overview

Hypernasality is one of the speech disorders in children with cleft palate (CP). It is mainly because of the nasalization of vowels, and that happens due to the addition of nasal formant and antiformant pairs in the vowel spectrum. The selection of the right medical treatment for hypernasality is based on the assessment of speech produced by CP children. The hypernasality assessment is done to find the severity of nasality present in their speech. In the clinical environment, the hypernasality assessment is done perceptually and the perceptual judgement is supplemented with the measurements obtained using instrumental methods. However, researchers are attempting hypernasality assessment based on the spectral analysis of hypernasal vowels present in the CP speech. In this technique, hypernasality severity detection is done using the features capable of capturing the spectral deviation in hypernasal vowel compared to normal vowel. The spectral characteristics of hypernasal vowels deviate from the normal due to the presence of additional nasal formant-antiformant pairs in the spectrum. The aim of this thesis is to develop an objective method based on the spectral analysis of speech for the assessment of hypernasality in CP speech. This has been done by exploring the temporal, sinusoidal model-based and cepstral features for capturing the spectral deviation in hypernasal vowels and using them for hypernasality severity detection. The temporal features capture the effect of spectral deviation by analyzing the speech and linear prediction (LP) residual signals. The sinusoidal model-based features capture spectral deviation by estimating the strength of low-frequency harmonics, whereas the cepstral features capture spectral deviation by modeling the envelope of the spectrum. In this thesis, features are first used for normal vs. hypernasal speech detection. Later the best performing features and the vowel space area (VSA) feature are used to propose a method for hypernasality severity detection. Finally, a MATLAB based graphical user interface (GUI) of the proposed method is also implemented for clinical application.

1.1 Introduction

Hypernasality is defined as excessive nasal resonance heard on vowels, and in the severe case on voiced consonants [1]. It is a speech disorder which reduces the intelligibility of speech [2]. Speakers with velopharyngeal dysfunction (VPD) produce hypernasal speech. VPD indicates a condition where the velopharyngeal valve does not close completely and consistently the velopharyngeal gap. VPD results in leakage of air through the nose during the production of oral sounds [3]. VPD is a general

term which is more specifically termed as follows [3]:

- **Velopharyngeal insufficiency**: This type of VPD is due to anatomical or structural defects, such as cleft palate. In this case, the movement of velum may be proper but its size is too short to close the velopharyngeal gap. It is the main reason for hypernasality in the speech of children with repaired CP.

- **Velopharyngeal incompetence**: This type of VPD is due to a neuromotor disorder, such as central nervous system damage (cerebral palsy or traumatic brain injury), or peripheral nervous system damage (Moebius Syndrome). In this case, the size of velum is normal but the movement of velopharyngeal structure is too poor to close the velopharyngeal gap. The hypernasality in the speech of individual with dysarthria and hearing impairment is due to this type of VPD.

- **Velopharyngeal mislearning**: This type of VPD is due to faulty development of appropriate articulation patterns. This VPD also sometimes causes the hypernasality in the speech of children with repaired CP.

Fig. 1.1, and Fig. 1.2 show the velopharyngeal insufficiency and velopharyngeal incompetence VPD respectively. These figures are taken from the book by author Ann W. Kummer [4]. The VPD causes the abnormal coupling of the nasal and oral cavities during the production of hypernasal speech. The hypernasal speech energy gets absorbed as it goes through the turbinates of the nasal cavity. Hence, the speech is often described as muffled or characterized by mumbling. All oral phonemes get nasalized in hypernasal speech due to the mixing of oral and nasal resonances, but hypernasality is particularly perceptible on vowels. This is because vowels are relatively long in duration and are typically not substituted with a different placement [3]. Among the vowels, hypernasality is more perceived on high vowels such as /i/, /u/ [1, 5] due to the high tongue position during their production [3]. The nasalization of voiced plosive results in substitution of these sounds by their nasal cognate (i.e., m/b, n/d, and ng/g) [3]. The other oral sounds (unvoiced plosive, fricatives, and affricates) can also be substituted because of nasalization. Hypernasality is observed in the speech of individuals with cleft palate [3], dysarthria [6] or hearing impairment [7]. It is a major speech disorder in children with CP and most of the children with CP are affected by this disorder. This thesis is particularly about the detection of hypernasality in CP speech. However, the features and the method proposed in this thesis can also be used for hypernasality detection in case of other disorders.

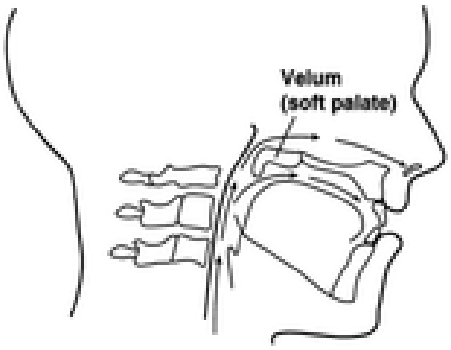

Figure 1.1: Velopharyngeal insufficiency. The figure is taken from the Ref. [4]

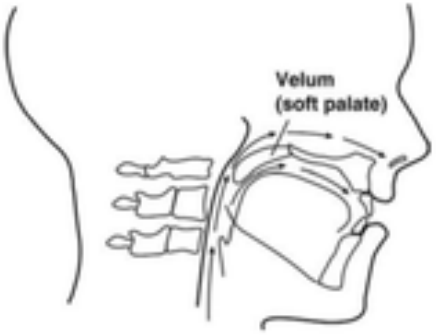

Figure 1.2: Velopharyngeal incompetence. The figure is taken from the Ref. [4]

1.2 Hypernasality in cleft palate speech

The cleft lip and palate (CLP) is the most common craniofacial birth defect in children. Worldwide it occurs in about 1 to 2 per 1000 birth. CLP constitutes almost two-thirds of the major facial defects and almost 80% of all orofacial clefts [8]. The cleft may be unilateral or bilateral and can be categorized into cleft lip (CL), cleft palate (CP) or CLP. There is also a subset of CP, which is hard to observe, called submucous CP. The action of the mouth and lips get affected in CLP due to the cleft. So the children born with CLP might have feeding problems, dental problems, middle ear fluid buildup and hearing loss and speech disorders associated with them. The speech disorders are a functional problem, which is mainly concerned with the CP rather than the CL [3]. Hypernasality, hyponasality, nasal air emission and/or turbulence, consonant production error, and voice disorder are five universally accepted disorders associated with the CP speech [1].

Hypernasality in CP speech is due to leakage of air into the nasal tract through the cleft in the hard and soft palate (velum). Hence first step taken as a part of hypernasality treatment is to repair the clefts by primary or sometimes secondary surgery. However, hypernasality persists in the speech of children with repaired CP. It happens because the velum gets short after surgery due to the stitching of the muscles. The short velum thus becomes insufficient to completely close the velopharyngeal gap during the production of voiced sounds. This phenomenon is called velopharyngeal insufficiency [3]. Choosing the right medical treatment for hypernasality in children with repaired CP depends on the speech assessment done by speech-language pathologists (SLPs) to find the severity of hypernasality. A variety of scales like equal-appearing interval, visual analog, and direct magnitude estimation have

been used for the severity grading of hypernasality [9]. But ultimately the grading involves mapping of speech onto a 4-point scale in the increasing severity order from 0 through 3 [1]. The 0 point on the severity scale represents the speech which is within the normal limit of nasality heard in regional speech and there is no perceptual evidence of hypernasality. Point 1 represents mild hypernasal speech. The nasality exceeds the regional speech nasality limit and heard mainly on high vowels. Point 2 represents moderate hypernasal speech. The nasality is perceived as pervasive in moderate hypernasal speech and it is heard on both high and low vowels. Finally, Point 3 represents the severe hypernasal speech having pervasive nasality and reduced understandability. The nasality is heard on all vowels and sometimes on voiced consonants also. The identity of some vowels may also be lost due to the severe hypernasality.

1.3 Hypernasality Assessment in cleft palate speech

The assessment of hypernasality is done to find the severity of nasality present in the speech. In the clinical environment, perceptual evaluation supplemented with velopharyngeal activity measurement using instrumental methods is used for the hypernasality assessment [10]. The clinical assessment is done at the single-word level as well as at the sentence level. For that, the test words in the consonant-vowel (CV) or consonant-vowel-consonant structure having voiced or unvoiced consonants and only one vowel (low or high) are considered. However, the words with high vowels only are suitable for mild hypernasality evaluation. The considered sentence should contain only one consonant and all vowels relevant to the language. Also, the test words or sentences should not contain nasal consonants. Some examples of test words and sentences are: *pea, tea, key, dad, coke, /Buy baby a bib/*, and */Bob is a baby boy/* [1]. The perceptual evaluation is performed by experienced SLPs by listening to the test words or sentences. Although the perceptual evaluation is considered as a gold standard in hypernasality assessment, it has issues that reduce their reliability.

The instrumental evaluation techniques can be categorized into two groups: direct and indirect [10]. The direct techniques such as X-Ray (Cephalometry) [11], videofluoroscopy [12], and nasendoscopy [13] are used to observe insufficient velopharyngeal port movement during the production of hypernasal speech. The severity of hypernasality is judged based on the amount of velopharyngeal gap size observed [14]. As the gap size increases, the severity varies from normal to severe hypernasality. The indirect techniques are used to study the aerodynamic and acoustic consequences of insufficient

velopharyngeal port movement. Unlike the direct techniques, indirect methods provide quantitative measure about the severity of hypernasality. The aerodynamic measurements are done to measure airflow and air pressure. These measurements are based on the principle that insufficient closure of the velopharyngeal gap causes the increase in nasal air escape which affects the nasal airflow and pressure [10, 15]. The mirror-fogging test [16] and the devices such as aerophonoscope [17], pneumotachograph [18], and warm-wire anemometer [19] are used to measure the nasal and oral airflow. The nasal airflow measurement is further used to develop a pressure-flow technique which objectively evaluates the velopharyngeal gap size [20].

The assessment of hypernasality is also done based on the acoustic measurements obtained from the speech produced by the children with CP. The acoustic measurements indicate nasality based on the relation between nasal and oral acoustic energy. The measurements are done with the accelerometry and nasometry techniques. In the accelerometry technique, accelerometers are placed on the outer surface of the nose and throat of the CP children during the utterance of test words. Accelerometers capture the nasal and oral accelerometer signals based on which indices such as Horri Oral Nasal coupling Index [21] and Nasal accelerometer vibration Index [22] are computed for the hypernasality measurement. In the nasometry technique, the relation between nasal and oral sound energy is computed with the Nasometer device in terms of The Oral Nasal Acoustic Ratio (TONAR) measure for the hypernasality measurement. The TONAR measure gives the percentage of "Nasalance", i.e. the ratio of nasal acoustic energy to the total oral-plus-nasal acoustic energy. Speech having the nasalance value above than a cutoff value is considered as the hypernasal speech. The cutoff value of nasalance is defined based on the values obtained for the normal speech. Nasometry is widely used for the clinical purpose, but accelerometry is rarely used because commercially it is not available as a package [10].

1.4 Issues with current hypernasality assessment techniques

The issues with the perceptual evaluation of hypernasality are that the perceptual judgement may vary among the SLPs. The variation is possible because of some factors such as abnormalities in pitch, loudness, voice quality and/or articulation which occurs in conjunction with the hypernasality [23], and these abnormalities may affect the perception of nasality in hypernasal speech [7, 24]. The variation in perceptual judgement may result in intra-rater and inter-rater disagreement. The reliability

of judgement also depends on the knowledge, experience, and mood of the SLPs. The perceptual evaluation also requires expert SLPs who's numbers are significantly less compared to the number of children with hypernasality. Issues are also associated with the instrumental evaluation techniques which are used to supplement the perceptual judgement. The direct techniques like nasendoscopy is an invasive method which may be harmful or painful to the children, and the videofluoroscopy may has ionization radiation effect [25]. The indirect techniques are radiation free, non-invasive, cause no discomfort, and are completely safe for the children. Hence, these techniques are more feasible to conduct large-scale studies, including more children and possibly more languages in the data sample. However, these techniques require more caution during the experiment because there can be some changes in aerodynamic or acoustic data that may not reflect the changes in velopharyngeal opening and nasal coupling. Further, the techniques require additional sensing device at the nose besides the microphone at the mouth for the measurement of hypernasality evidence. The issues with the Nasometer device is that it cannot be used for the prerecorded speech [26].

1.5 Hypernasality assessment based on spectral analysis

The researchers are working on another approach for hypernasality assessment based on the spectral analysis of speech using a digital signal processing techniques. In this approach, features are extracted from vowels present in hypernasal speech to capture their specific spectral characteristics. The features thus obtained are used for the hypernasality severity detection based on some threshold or using a classifier. The vowels present in hypernasal speech get nasalized due to the coupling of the nasal tract with the oral tract. The coupling adds the nasal pole-zero pairs in nasalized vowels [27]. The addition always happens in pairs because the nasal tract is coupled as a shunt side branch, and according to circuit theory, any zero due to the shunt side branch is always paired with a pole [27]. The presence of nasal pole-zero pairs affects the spectral characteristics of hypernasal vowels. The spectral analysis shows these effects in the form of presence of nasal formant and antiformant pairs, mainly in the vicinity of first formant (F_1) [26,28], reduction in strength of F_1 [29] and flattening of spectrum [30] which are regarded as cues for nasality. However, due to the complex interaction of resonances of the nasal and oral tract and substantial difference in the shape of the nasal tract across the speaker, it has been difficult to arrive at a single reliable spectral measure of nasalization. Nonetheless, with an understanding of the basic principles involved in nasalization of vowels, researchers have proposed

various features capable of capturing the spectral cues for hypernasality detection and also for the hypernasality severity detection. The hypernasality detection is done by normal vs. hypernasal speech classification, whereas hypernasality severity detection is done by multi-class (normal, mild, moderate and severe) classification.

Feature-based on Teager energy operator (TEO) profile along with Mel-frequency cepstral coefficient (MFCC) feature [31, 32], linear prediction cepstral coefficients (LPCC) feature [33], acoustic, noise and cepstral analysis based features [34], and non-linear dynamics features along with entropy features [35,36], are used to capture the different cues of nasality for hypernasality detection. Further, the feature extracted from a high spectral resolution group delay spectrum [26], features based on the distribution of energy [37, 38] are also used for hypernasality detection. The database used in all the above works consists of vowel phonemes. Recently, hypernasality detection work is done using recorded sentences speech database using jitter, shimmer, MFCC, bionic wavelet transform entropy and bionic wavelet transform energy features [39]. The above mentioned works on hypernasality detection give good accuracy ranging from the 70% to 90%. However, doctors and SLPs are more interested in the severity grading of hypernasal speech. This is because severity grading information can be related to the velopharyngeal gap size in children with CP. Even so, severity grading is highly important, very few works have been done in this regard. The relation between spectral characteristics and perceived hypernasality using one-third octave spectra [40], and hypernasality severity detection using formant feature [38], hypernasality severity analysis using vowel space area (VSA) feature [41] are few works which have been done related to severity grading of hypernasality.

Spectral analysis based technique of hypernasality assessment is non-invasive, radiation-free, objective, cheaper, and unbiased [26]. The technique requires mainly a good quality microphone and a computer for the recording and evaluation of the speech. It is comfortable to children and naturalness of speech remains maintained.

1.6 Motivation for the present work

The present work aims to explore various features for hypernasality detection and propose a method for hypernasality assessment based on the hypernasality severity detection using these features. As discussed in the previous section, various works have been done for hypernasality detection using different features. The spectral analysis of hypernasal vowels shows the presence of nasal formant-

antiformant pairs in the spectrum which centralize the energy of vowels in lower-frequencies. Hence, in some hypernasality detection works [31,37,38] features are extracted from the low-frequency region, below a predefined cutoff frequency just above the F_1, of the vowel spectrum. This has been done to capture the centralized energy in the low-frequency region of hypernasal vowels. But there is no definite way to find the proper cutoff frequency. Hence, the value of cutoff frequency varies from vowel to vowel and for a particular vowel also, the detection accuracy depends upon the cutoff frequency. The problem of finding appropriate cutoff frequency gives the motivation to explore a feature that can capture the centralized energy in the low-frequency region of hypernasal vowels without the need for a predefined cutoff frequency.

The previous study in [42] shows that the presence of nasal formants and antiformants in the hypernasal vowel spectrum causes the modification in the temporal characteristics of linear prediction (LP) residual signal by adding the undesirable signal components in the vicinity of each glottal closure instants (GCIs). So the temporal characteristics of the residual signal corresponding to the hypernasal vowel may deviate from the residual signal corresponding to normal vowel. However, the residual signal has not been explored for hypernasality detection yet. So, the temporal deviation in the residual signal of hypernasal speech motivates to extract the feature which can capture the nasality evidence from the residual signal.

The vowel spectrum consists of the fundamental frequency and its harmonics. The larynx is considered to function in nearly a similar way during the production of normal and hypernasal vowels [43]. Hence, harmonic frequency locations for normal and hypernasal vowels are expected to be similar. However, the addition of nasal formants and antiformants in hypernasal vowels, respectively, increases and decreases the strength of harmonics around the frequency of their addition. Hence, the pattern of the strength of harmonics in hypernasal vowels may differ compared to normal vowels. This gives the motivation to analyze the strength of harmonics in lower-frequencies and compute features based on harmonics strength for hypernasality detection.

Some other works on hypernasality detection in the literature [32, 35], use the MFCC feature to capture the spectral deviation in hypernasal speech compared to normal. The cepstral coefficients are computed from the magnitude spectrum using a window of size 20-30 ms to model the envelope of the spectrum. As the envelope of the spectrum represents the vocal tract system characteristics, so the MFCC feature captures the average characteristics over all the pitch periods within that window

segment. However, studies show that the vocal tract characteristics may change significantly even within a pitch period [44]. Also, the phase spectrum of the signal is ignored while computing the MFCC feature. This gives the motivation to explore the cepstral feature computed from the phase spectrum of the signal with the window size of less than or equal to the pitch period. Moreover, studies also show that for the high pitch children's speech, the MFCC feature has a high variance for higher coefficients [45]. The variance is due to the inability of filter-banks to sufficiently smooth out the pitch harmonic present in the magnitude spectrum. The insufficient smoothing in low-frequencies, where the nasality evidence is mainly present may affect the hypernasality detection accuracy for the MFCC feature. This gives the motivation to explore another cepstral feature that can smooth out the pitch harmonic effect present in the MFCC feature. Also, the researchers have explored spectral moment based frequency features as an alternative to the MFCC feature for speech recognition [46], which gives the motivation to explore these features for hypernasality detection.

Almost all the works on hypernasality are done for normal vs hypernasal speech classification. But doctors and SLPs are more concerned about the severity grading of hypernasality. Also, the classification is done individually on different vowels and their accuracies are different. To the best of our knowledge, no attempt has been made in the literature to detect hypernasality using a feature that is extracted simultaneously from different vowels. This gives the motivation to investigate hypernasality severity detection using the feature extracted simultaneously from different vowels.

The major contributions of the current thesis are as follows:

- An isolated word-level speech database is developed for hypernasality detection.
- Temporal features are used for hypernasality detection which does not require a predefined cutoff frequency for capturing the nasality evidence.
- Strength of harmonics in the low-frequency region and their frequencies computed using the sinusoidal model of speech are explored for hypernasality detection.
- The cepstral coefficients computed from the group delay spectrum, cepstral smooth spectrum obtained after pitch-adaptive liftering and cepstral coefficients augmented with the spectral moment based frequency features are explored for hypernasality detection.
- The hypernasality severity detection is attempted using the nasality score between [0 to 1] for the severity grading of normal, mild and moderate-severe hypernasal speech.

- A MATLAB based GUI of the proposed hypernasality severity detection method is implemented for the clinical application.

1.7 Organization of the thesis

The thesis is organized as follows.

- Chapter 2 firstly, reviews the spectral characteristics of nasalized vowels, proposed nasality cues for nasalized vowels and works done for normal and nasalized vowel detection. The chapter later discuses the analysis of hypernasal speech and works done for detection and severity grading of hypernasality using the features capable of capturing the nasality cues.

- Chapter 3 explores two temporal features, vocal tract constriction (VTC) and peak to side-lobe ratio (PSR) for hypernasality detection. Temporal features are the time-domain processing based features in this thesis. The VTC feature is extracted from the speech signal to capture the prominence of lower-frequencies. The PSR feature is extracted from the residual signal of speech to capture the characteristics of residual signal in the vicinity of each epoch location. Both the features capture the different aspects of nasality evidence without involving any low-pass filtering. The performance of combined (VTC+PSR) feature for hypernasality detection is evaluated on the database containing hypernasal and normal vowels and it is compared with the performance obtained for baseline features.

- Chapter 4 explores the normalized harmonics amplitude (NHA), harmonics amplitude ratio (HAR) and prominent harmonics frequency (PHF) features for hypernasality detection. The features are based on the strength of harmonics in lower frequencies computed using the sinusoidal model of speech. The performance of combined (NHA+HAR+DHF) feature is compared with baseline features and MFCC features.

- Chapter 5 explores three cepstral features namely, Hilbert envelope of numerator of group delay function (HNGDF), pitch-adaptive Mel-frequency cepstral coefficients (PAMFCC), and spectral moment features augmented with low-order cepstral coefficients (SMAC) individually for hypernasality detection. The HNGDF feature is computed from the group delay spectrum, the PAMFCC feature is extracted from the cepstral smooth spectrum obtained after pitch-adaptive

liftering, and the SMAC feature is the spectral moment based frequency feature which is augmented with the low order cepstral coefficients. The performances of these three features are compared with the MFCC feature. Finally, the chapter compares the performance of temporal, sinusoidal model-based features, cepstral features and inter combination of these features for hypernasality detection.

- In Chapter 6, a method for hypernasality severity detection is proposed. The method is based on the nasality score between [0 to 1] corresponding to the speaker's speech. The method is also validated by estimating the nasality score corresponding to the speech of children with CP. Finally, a MATLAB based graphical user interface GUI of the proposed method is implemented for the clinical application.

- In Chapter 7, a summary of the present work is reported by discussing the contributions made in this thesis. Finally, the chapter discusses the directions for future work considering the issues related to this thesis.

2

Objective methods of hypernasality assessment in CP speech: A review

Contents

Overview

The instrumental methods of hypernasality assessment, which are objective by nature, are used to supplement the perceptual rating. This is done because the perceptual assessment is a subjective method and its rating may vary among the SLPs. However, the instrumental method also has limitations. Hence, researchers have been working to develop a new objective hypernasality assessment method based on spectral analysis of CP speech using digital signal processing techniques. More precisely, in this method nasality cues for the nasalized vowel are captured by various features and hypernasality severity detection is performed using these features for hypernasality assessment. The features may be temporal or based on formant analysis or cepstral coefficients. This chapter provides a review of various hypernasality detection works done based on features for hypernasality assessment. The chapter first reviews the various nasality cues proposed in the literature for nasalized vowels. Later, works done for the normal vs. hypernasal speech detection and hypernasality severity detection are reviewed. In the last, a review of existing clinical methods used for hypernasality assessment is also discussed.

2.1 Introduction

The assessment of hypernasal speech is needed for deciding the appropriate treatment of hypernasality disorder. In the clinical environment, as mentioned in the previous chapter, the assessment is done using the perceptual rating given by the experienced SLPs. As perceptual evaluation is subjective by nature, so its rating may vary among the SLPs. Hence, the perceptual rating is supplemented with the objective measures obtained from direct and indirect instrumental methods of hypernasality assessment [10]. The direct methods are used for observing the velopharyngeal gap size and the change in acoustic characteristics due to the gap is measured by the indirect method. However, radiation hazard and discomfort due to the invasive instrument are the examples of some limitations which are always associated with the instrumental technique. Considering the problems associated with the perceptual and instrumental method of hypernasality assessment, researchers have been working to develop another objective method based on spectral analysis of speech using digital signal processing. The technique is simple as it mainly requires a microphone and a computer to compute nasality measure. Also, the technique is non-invasive and cost-effective [26].

Fig. 2.1 shows the block diagram representing the steps involve in the spectral analysis based method for hypernasality assessment. In this method, at the first stage, the vowel region is detected

Figure 2.1: Block diagram showing the steps involved in spectral analysis of speech-based hypernasality detection.

from the speech stimuli uttered by the children with CP either by manual marking or by automatic detection method. In the second stage, various features are extracted from the vowel regions through short-term speech processing, which can capture the nasality cues. At the final stage, these features are used for hypernasality severity detection using a classifier or regression model. In the literature, researchers have mainly focused on normal vs. hypernasal speech detection [26, 31–38], and very few works have been done for hypernasality severity detection [38, 40]. The features which have been used in these works can be categorized into temporal features (time-domain processing based features), features based on oral and nasal formants strength, cepstral features and inter combination of these features. The temporal features are extracted from the speech signal in time domain to capture the centralized energy in lower frequencies [38]. The features based on formant analysis are extracted from the short-term speech spectrum after detecting nasal and oral formant locations and strengths. The cepstral features are also extracted from the speech spectrum by modeling the spectral envelope with the cepstral coefficients. The combined features are the inter combination of temporal features, formants based features and cepstral features. These categories of features are motivated from the fact that a nasalized vowel spectrum contains nasal formant and antiformant pairs in hypernasal vowel spectrum which enhance the energy in lower frequencies, affect the strength of oral formants, especially F_1, causes the shift in formant locations and affect the envelope of the spectrum [26, 37, 38]. However, extracting these features may have some issues for the case of children speech. To compute energy or the formant strength and location from the lower-frequencies, speech is typically low pass filtered with a predefined cutoff frequency. The variation in cutoff frequency may affect the hypernasality detection accuracy. So finding appropriate cutoff frequency may be an issue. Also, the detection of the closely spaced nasal and oral formants in high pitch children speech may not be accurate. Moreover, the high pitch also affects the spectral envelope in children speech. The objective of this thesis is to explore new features that can be used to develop a method for hypernasality assessment based on the hypernasality severity grading. Therefore, a detailed review of all the works done for hypernasality detection and

severity grading is needed to know about the database and the features along with their advantages and disadvantages. As the features used for hypernasality detection are based on the nasality cues proposed for nasalized vowels and similar to features used for normal vs. nasalized vowel detection, the review of nasality cues and works done for normal vs. nasalized vowel detection are also needed to understand the characteristics of hypernasal vowels. So in this chapter, firstly spectral characteristics of nasalized vowel and works done for normal vs. nasalized vowel classification are reviewed. Later, works done using temporal features, formant analysis based features and cepstral features for normal vs. hypernasal speech detection and hypernasality severity detection are reviewed. In the last, existing clinical methods used for hypernasality assessment has been also reviewed so that a practical system perspective of signal processing based method of hypernasality assessment can be better understand

The rest of the chapter is organized as follows: Section 2.2 discuses the works done to describe the nasality cues for nasalized vowel and works done for normal vs. nasalized vowel detection. Section 2.3 discuses the analysis of hypernasal speech. Section 2.4 and Section 2.5 respectively discuses the works done for hypernasality detection and severity grading using temporal features, formant strength features and cepstral features. Section 2.6 describes the existing clinical methods used for hypernasality assessment. Section 2.7 presents the summary of all the works done for hypernasality detection and the scope of the present work.

2.2 Spectral characteristics and detection of nasalized vowel

2.2.1 Spectral characteristics of nasalized vowel

The nasalization of vowels can be broadly divided into three categories: 1) coarticulatory nasalization, 2) phonemic nasalization, and 3) functional nasalization. Coarticulatory nasalization occurs when nasals sounds /m/, /n/ or / ŋ/ comes adjacent to vowels as in /mama/ and /mimi/ words. Phonemic nasalization is the category where language has some vowels which are nasalized such as /ə̃/ in the Hindi language. Functional nasalization is due to defects in the functionality of the velopharyngeal mechanism as in the case of CP. The reason for the nasalization of vowels in all these three categories is the coupling of the nasal tract with the oral tract due to incomplete closing of the velopharyngeal gap. The nasal tract is a complicated structure. It has two parallel passages that end with two nostrils. The areas of two passages are different which makes the passages asymmetric [47, 48]. The nasal tract has four major paranasal cavities which are called sinuses. These sinuses are: maxillary

sinus, frontal sinus, sphenoidal sinus, and ethmoidal sinus. Among these sinuses, maxillary sinus has the largest volume, whereas the frontal sinus has the smallest volume [49]. Unlike the oral tract, the nasal tract is a static tract and does not have any muscles which can dynamically vary its shape. However, the velum can cause some amount of dynamic variation by changing its position. Due to the coupling of the nasal tract, the shunt branch tube model of speech production mechanisms is used to study the acoustic properties of the nasalized vowel. The oral tract is considered as the main tube and the nasal tract as the shunt branch tube. The nasal tract introduces several pole-zero pairs in the nasalized vowel speech and these pole-zero pairs appear as additional formant and antiformant pairs in the nasalized vowel spectrum.

The nasalized vowels have been studied in the literature to define the nasality cues. Delattre [50] characterized the role of different formants in nasality perception and suggested that the main nasality characteristics lie in the low-frequency region. The author showed that the lowering in amplitude and widening of the F_1 can considerably nasalize the oral vowel. The study also proposed two fixed peaks at 250 and 2000 Hz and a variable peak at 900 Hz as the formants of nasal sounds. Hattori et al. [51] tried to find the effect of the nasalization on five Japanese vowels. The authors report 1) the presence of a dull formant around 250 Hz, 2) an antiformant around 500 Hz, and 3) comparatively weak and diffused peaks between the formants (mainly in the frequency range of 1000 to 2500 Hz) as nasality cues for nasalized vowels. The study also tried to find the acoustic correlates of these cues with the perception of nasality in the sound. It was reported that for open vowels such as /a/, the addition of a formant at the frequency of 250 Hz alone gives some perception of nasality, but the addition of an antiformant at 500 Hz alone does not give any nasality perception. However, the addition of formant and antiformant together gives an improved perception of nasality. For close vowel such as /i/ and /u/, it was reported that modification in the higher frequencies of spectrum, in terms of additional formants between the existing formants, is necessary for nasality perception. Fant [27] studied the acoustic characteristics of nasalized vowels and observed the presence of nasal pole-zero pairs in nasalized vowels. The study observed a reduction in F_1 strength, increase in bandwidth and frequency of F_1 in nasalized vowels compared to normal vowels. The study concluded that the acoustic characteristics of nasalized vowels depend on the vowel, speaker and the area of velopharyngeal coupling. The study done by Dickson [43] observes an increase in bandwidth of F_1 and second formant (F_2), an increase or decrease in the intensity of harmonics corresponding to F_1, F_2 and third formant (F_3) frequencies

as the cues for nasality. Carre et al. [52] studied the nasalization of vowels in CVCV words uttered by several speakers. The C in CVCV stands for the nasal consonants /m/, /n/, / ŋ/ and voiced stops (/b/,/d/,/g/), whereas the V stands for a vowel. The analysis of CVCV words is done using predictive coding (to detect formants) and cepstral prediction technique (to detect antiformants). The result shows that the nasalization of vowels such as /a/ and /æ/ is due to the weakness of F_1, whereas it is due to the presence of a nasal formant with large bandwidth between F_1 and the F_2 for vowel /i/.

Martin [53] proposed four cues for nasalized vowels. The first and the most important reported cue was the reduction in the intensity of the F_1. The reduction ranges from 7 dB for /i/ to 13 dB for /u/ vowel, and it is due to the damping characteristics of the nasal tract. The second reported cue was the presence of antiformants in the nasalized vowel spectrum. The location of antiformants is reported around 2400 Hz for /a/ vowel, 2100 Hz for /i/ vowel and 1400 Hz for /æ/ vowel. The third reported cue was the reinforced harmonics in the spectrum between the formants which are called nasal formants. It was also reported that the frequency locations and strength of nasal formants and antiformants depend upon the vowel, speaker and degree of nasal tract coupling. The fourth reported cue was the shift in the relative frequency positions of formants. The study also points out that whenever a vowel is perceived as nasalized, at least one of the four cues will be present in the spectrum. The study suggests that an objective measure for the clinical purpose can be developed based on the harmonics-intensity-measurement-procedure because the basic difference between normal and hypernasal vowels is found in the strength of harmonics. The sweep-tone measurement of the vocal tract done by Fujimura et al. [54] states that the nasalized vowels contain two kinds of formants: nasal formants which are paired with antiformant and the shifted oral formants. The frequency gap between the nasal formant and antiformants increases when the coupling between nasal tract and oral tract increases, and it decreases when the coupling decreases. All the formants, nasal or oral, shift upwards in the frequency domain as the degree of coupling increases. The authors suggested that the lowest frequency formant in the nasalized vowel can be a nasal formant or a shifted oral formant, and this decision depends upon a critical frequency. If the normal vowel has the F_1 at a higher frequency than the critical frequency, the lowest frequency formant will be a nasal formant. On the other hand, if the F_1 lies in the lower-frequency, the lowest frequency formant will be an oral formant. The critical frequency was defined as the lowest resonance frequency of the nasal tract closed at the coupling end. The sweep-tone measurement was done for nasalized back vowels /u/, /o/ and /ɑ/. For the case of

/u/ vowel, the author observed first peak at a slightly lower frequency than the original F_1 along with a pair of a valley and a peak above the first peak. For /o/ and /ɑ/ vowels, the antiformant occurs at the same frequency as in /u/ vowel, but the nasal formant in /o/ vowel is slightly higher than the /u/ vowel and it is at still higher frequency in /ɑ/ vowel. The study by Maeda [25, 30] suggests that the main cue for the nasalization of vowels is the flattening of the spectrum in the frequency range of 300 to 2500 Hz. In some other studies, [55–57], the presence of nasal formant below the F_1 due to the sinus cavities has been suggested as another important nasality cue.

Hawkins et al. [28, 58] synthesised five nasalized vowels /i/, /e/, /ɑ/, /o/ and /u/ by inserting a pole-zero pair in the vicinity of F_1 to judge the perception of nasality in these vowels. The stimuli were perceived by people from different language backgrounds and it was postulated that there will always be an acoustic cue perceived as a nasal, independent of the vowels, regardless of the language background of the listeners. It was also postulated that there are one or more cues that have a different degree of nasality perception in different languages. The experiment proposed the degree of prominence of the spectral peak in the vicinity of the F_1 as a basic acoustic cues for vowel nasalization. The experiment also proposed a shift in the center of gravity of the low-frequency spectral prominence and changes in overall spectral balance as the secondary cues in nasalized vowels. Bognar et al. [59] reported the formant shifts and introduction of two pole-zero pairs, one below F_1, the other between F_2 and F_3 as the main acoustic cues for vowel nasalization. In the study, phonemic and phonetic perceptual tests on the synthesized normal and nasal /ɛ/ vowel shows that the formant shifts and pole-zero separation contribute almost equally for the phonemic identification, but pole-zero separation has a comparatively stronger effect in the phonetic judgement on nasality. Beddor et al. [60] proposed that a well-defined spectral peak and the overall spectral envelope of the low-frequency region, consisting of F_1, nasal formant and F_2 of the nasalized vowel, both contribute to the perception of nasality. Arai [61] studied the acoustic characteristics of nasalized vowels to investigate whether the shift in formants in nasalized vowels happens and is the human perception compensates for such shifts. The study found that nasal tract coupling modifies the vowel spectrum and shifting of formants happens. The measurement of formant frequencies of various nasalized vowels showed that the F_1 tends to shift in a more central direction. The study founds that listeners were perceiving the vowels in the English language as the same phoneme regardless of nasalization. In other words, the study confirmed that listeners have the capability to compensates for formant shift in English language vowels. Further, the

author performed the perceptual experiment to examine the compensation effect on synthetic speech. For that, the author synthesized a non-nasal and a nasal vowel with the same oral formant frequency separately and made a continuum using two vowels. The result of the perceptual experiment showed that the nasal vowels are more correctly identified than the non-nasal vowels, which supports the existence of the compensation effect.

Chen [62,63] studied the acoustic characteristics of hearing-impaired children speech and nasalized vowels present in normal-hearing adults speech. The study suggests an extra pole-zero pair between the F_1 and F_2 and the reduced F_1 prominence as cues for the nasalized vowels. The study also proposed a parameter $(A1-P1)$, which is a difference between the amplitude of F_1 and the amplitude of the extra peak between F_1 and F_2, to measure the nasality. Further, it was found that the measure correlates with listener judgements of the degree of vowel nasality in utterances of hearing-impaired and normal-hearing children. In another study [64], Chen proposed a new parameter $(A1-P0)$, which is a difference between the amplitude of F_1 and the amplitude of the peak below the F_1, to measure the nasality in nasalized vowels. Styler [65] examined 22 existing nasality cues for nasalized vowels to find the optimal acoustic nasality cue on the database of 4778 oral and nasal vowels in English and French languages. The study shows that the cues $A1-P0$, F_1 bandwidth and spectral tilt are more promising cues among the 22 cues. However, these cues, particularly $A1-P0$, show variation across speakers and vowels within each language. The finding of study suggests that the acoustic characteristics of nasalized vowels are language as well as the speaker-specific and require speaker normalization for across-speaker comparison. Also, these cues can not be considered constant across different languages. Paul et al. [66] studied the glottal volume velocity waveform of nasalized vowels obtained from inverse LP analysis of speech. The glottographic analysis was performed on non-nasal and nasal sounds synthesized using a transmission line model of the vocal and nasal tracts. The analysis was also performed on nasal and non-nasal vowel sounds produced by a male speaker. The study shows that nasalization introduces extraneous ripple onto the closed phase portion of the glottal volume velocity waveform, but there is no change in the observed open quotient. The study states that the presence of ripple is due to a closely spaced antiresonance pair that is not compensated by the all-zero LP inverse filter.

Some simulation studies are also done in the literature which confirms the nasality cues in nasalized vowels. In one of the first simulation study of nasalized vowels derived from electrical analogs of

anatomical structures, House et al. [29] reported a differential reduction in amplitude of F_1, increase in formants bandwidth and an upward shift in formants frequency in nasalized vowels compared to normal vowels. Further, a reduction in the overall level of the vowel, introduction of additional formants and antiformants, elimination of the F_3 and inconsistency in higher formants were also reported in this investigation. Based on these cues it was observed that when the effect of these cues reaches to an appropriate level, the nasality is perceived in the vowels. Feng et al. [67] characterized the acoustic properties of nasalized vowels by considering vowels as a dynamic trend from an oral configuration toward an ŋ-like configuration. The author proposed ŋ-like configuration, which corresponds to the pharyngonasal tract having first two resonance frequencies at about 300 and 1000 Hz, as a target for nasalized vowels. This consideration allows explaining the characteristics of vowels between two simple configurations corresponding to oral vowel and ŋ-like configuration. The authors presented the pole-zero characteristics of 11 French vowels using this consideration through simulation. Nguyen [68] studied the effects of nasal tract coupling on the vowels and proposed that new formants appear in the vowel spectrum when the nasal tract gets connected with the oral tract. The author computed the variation in input impedance with frequency by looking from the coupling point of nasal tract into the three tubes. The computation is done with a numerical harmonic simulation of the transmission-line analogy taking into radiation, heat, viscosity and wall vibration losses into account. With these data, the study defines the nasal formant at the frequency which is obtained by looking into the nasal tract from the coupling point where reactance is infinite and nasal antiformant at the frequency where reactance is zero. The interaction of three branches gives the two newly created formants by splitting the original formant by nasal antiformant. One of these formants is termed as the primary formant and it is the original formant with a shifted frequency. The other formant is termed as the secondary formant and it is additionally created formant. The nasal antiformant frequencies are greatly diminished in the spectrum and it is due to loss of energy in the nasal tract. Rong et al. [69] studied the acoustic characteristics of nasalized vowels considering the effects of velopharyngeal opening and oral articulation. The study is done using the vocal tract area functions of an American English speaker. The work was done to study the spectral evolutions of nasalization of three English vowels /a/, /i/, and /u/ by simulating transfer functions for vowels with only velar movement, and for different nasal consonant-vowel utterances, which include both velar and oral movements. The study finds that if oral articulation and velar movement are coordinated to attenuate the nasal acoustic

features, then compensatory articulation can be developed to reduce nasality. This can be done either by adjusting the articulatory placement for isolated nasalized vowels or by changing the relative timing of coarticulatory movements during speech production. The results of study also demonstrate the effect of oral articulation on the acoustic characteristics of nasalized vowels. Pruthi et al. [69] simulated and analyzed the acoustics of nasalized vowel using the vocal tract area functions of an American English speaker, recorded using magnetic resonance imaging. The work studied the velar coupling area, asymmetry of nasal passages and the sinus cavities which are the three most important sources of acoustic variability in the production of nasalized vowels using the vocal tract models and susceptance plots. The simulation results of two nasal passages show the addition of extra pole-zero pairs due to the asymmetry between the nasal passages. When maxillary and sphenoidal sinuses are included in the simulation, the result shows that each sinus can add one pole-zero pair in the spectrum. Further, the result also shows that the right maxillary sinus adds a pole-zero pair at the lowest frequency.

In summary, the above-mentioned works for the characterization of nasalized vowels proposed the following acoustic and perceptual cues:

- Reduction in the amplitude of F_1 and increment in its bandwidth.
- Upward shift in frequency of F_1.
- Changes in the amplitude and frequency location of F_2 and F_3.
- Reduction in the overall level of the vowel.
- Flattening of the spectrum in the frequency range of 300 to 2500 Hz.
- Addition of nasal pole-zero pairs in the spectrum, mainly in the vicinity of F_1. The location and strength of these pole-zero pairs depend upon the vowel, speaker and degree of nasal tract coupling.
- Shifts in the center of gravity of the low-frequency spectral prominence.

2.2.2 Detection of nasalized vowels

The nasality cues discussed in the previous subsection are used to develop features for the classification of normal vs. nasalized vowels. The detection of nasalized vowels is useful for automatic

speech recognition. This is because the nasalized vowels give information about the presence of short or almost absent nasal murmur such as in words like /smack/ or /can't/. It is also important for correct formant tracking. In literature, attempts have been made for nasalized vowel detection. Glass et al. [70] attempted nasalized vowel detection using six features: center of mass, standard deviation, maximum resonance percentage, minimum resonance percentage, maximum resonance dip, and minimum resonance difference. The work was done using 685 samples of nasalized vowels and 500 samples of non-nasalized vowels. The experiment is done on the leave one speaker out basis and detection accuracy of 74% is reported. Pruthi and Carol [71] proposed the nine automatically extractable acoustic parameters for nasalized vowel detection. The parameters are based on extra peaks at low frequencies and relative amplitude of these peaks compared to the F_1, extra peaks in the whole spectrum, reduction of the strength of F_1, increase in bandwidth of F_1 and spectral flattening at lower frequencies. The performance of these parameters was tested on several databases with different sampling rates and recording conditions. The work reports the classification accuracies of 96.28%, 77.90% and 69.58% on StoryDB, TIMIT, and WS96/97 databases, respectively using the SVM classifier. Najnin et al. [72] performed the nasalized vowel detection based on the cepstral features derived from the product spectrum which is a phase spectrum. The feature is termed as the Mel-frequency product spectrum cepstral coefficients. The feature is fed to a linear discriminant analysis based classifier for the detection. The experiment was done using the TIMIT database and the result shows that the features outperform the state-of-the-art features in the task of detecting nasalized vowels in clean as well as different noisy conditions.

2.3 Analysis of hypernasal speech

The above section discusses the spectral characteristics of nasalized vowels proposed in the literature and the works done for the detection of nasalized vowels. As the vowels in hypernasal speech are also nasalized, so researchers have analyzed the hypernasal vowels to characterize the speech. In this section, various studies done based on the temporal and spectral analysis of hypernasal speech is presented.

Ha et al. [73] compared the temporal characteristics of nasalization in CP and normal speakers to find the relation between temporal measures and perceived nasality. The database used for the study was collected from 15 speakers with CP speech and 15 speakers having normal speech. The

sound coming from mouth and nose during the production of /pamap/, /pimip/, and /pumup/ words are recorded for the comparison. The result shows that speakers with CP have a longer duration of nasalization than speakers without CP and the nasalization grew longer as the degree of perceived hypernasality increased. Further, the speakers without CP showed larger nasalization-duration ratios in the high vowel contexts than in the low vowel context, however, differences in nasalization-duration ratios among the vowel contexts are not observed in speakers with CP. Kozaki-Yamaguchi et al. [74] worked to find the relation between perception of hypernasality and its physical correlation by performing three experiments on Japanese /i/ vowel. The first experiment was a spectral analysis of vowels obtained from five CP patients and six velum resection patients to find the spectral features related to hypernasality. The second experiment was conducted using various spectrally modified vowels to find the relation between spectral feature and auditory perception. Finally, the third experiment was done to analyze the estimated spectral envelopes to clarify the relation between spectral feature and velopharyngeal opening. The first and second experiment proposed the broadening of $F1$ bandwidth, an additional peak at around 1 kHz [$P1$], a decrease in the magnitude of $F2$ and dip between $F2$ and $F3$ [$D2$] as the four important cues for hypernasality in the CP speech. The second experiment further emphasized that the simultaneous modification of the decreases in $F2$ magnitude and the presence of $D2$ are very important cues for the auditory perception of hypernasality. The result of the third experiment showed that the presence of dips, due to the coupling of nasal tract, causes spectral modification in hypernasal vowels. The experiment also showed that the dip regions spread from the low-frequency region around $F1$ to the high-frequency region above $F2$ as the velopharyngeal gap increases.

Eshghi et al. [75] proposed a new discriminating cue, the magnitude of rising and falling slope amplitudes from the vowel signal, for hypernasality detection in CP speech. The database used for the analysis consisted of two words /iti/ and /iki/ produced by two normal children and two CP children. The extraction of vowel from the words was done using PRAAT software [76]. The result showed that the mean falling and rising slopes of the amplitude in the nasalized vowel are smaller than those of the oral vowel. Vikram et al. [77] enhanced the CP speech by suppressing the nasal formant and enhancing the spectral peak-valley. The study pointed out that low-frequency nasal formant is found around 250 Hz in /a/ vowel and around 1000 Hz in /i/ and /u/ vowels. The study also points out that the peak-to-valley ratio in hypernasal vowels gets affected due to the presence of

these nasal formants. Nikitha et al. [41] analyzed the normal, mild and moderate-severe hypernasal speech using the vowel space area (VSA). The VSA is computed from three vowels /a/, /i/ and /u/ in sustained phonation and from the vowels in the context of /p/, /t/, and /k/ consonants. The analysis showed that the VSA reduces for the hypernasal speech compared to normal speech. For the CP speech, the VSA is more for the vowels in the context of /p/, followed by /t/, and lastly by /k/. Statistical analysis yielded the significant difference in VSA value ($p < 0.05$) for normal, mild and moderate-severe hypernasal speech.

2.4 Detection of hypernasal speech

In the previous section, works done for the analysis of hypernasal vowels in CP speech were discussed. The analysis works show similar nasality cues as for the coarticulatory nasalized vowels. So researchers have proposed features to capture nasality cues for the detection and severity grading of hypernasal speech. The features can be classified into temporal features, features based on formant analysis, cepstral features or inter combination of these features. In this section, the works done for hypernasality detection are discussed. The works which are done for hypernasality severity detection are discussed in the next section.

2.4.1 Hypernasality detection using temporal features

The nasal formant and antiformant pairs get added in the spectrum of nasalized vowels present in hypernasal CP speech and affect the spectrum compared to normal speech. Generally spectral features are used to capture the deviation in the spectrum. However, the effect of nasal formant and antiformant pairs can also be captured by the features computed from the speech signal in the time domain. The first work done for hypernasality detection using temporal feature was based on the TEO profile of the speech signal. The hypernasal speech is a multi-component signal because it contains pole-zero pairs corresponding to the nasal tract besides the poles corresponding to the oral tract. Hence, Cairns et al. [31, 78] used a nonlinear TEO which is defined as [79],

$$\psi_d[x(n)] = x^2(n) - x(n-1)x(n+1), \tag{2.1}$$

for the detection of hypernasal speech. The operator is very much sensitive towards the multi-components signal. The difference between Teager energy operator (TEO) profile for low pass and the band pass filtered signal was proposed as a feature for hypernasality detection and it is quantified in

the form of correlation coefficient given by,

$$r = \frac{C}{\sigma_{LPF} \times \sigma_{BPF}}, \tag{2.2}$$

where, σ_{LPF} is the TEO profile of low-pass filtered signal and σ_{BPF} is the TEO profile of high-pass filtered signal. The low pass filtering of speech is done with the cutoff frequency just above the F_1 and the band pass filtering was done around the F_1. Both low pass and band pass filtered signals of normal speech vowels contains their F_1 only, whereas the low pass filtered signal of hypernasal speech vowels contain their F_1 and extra nasal formant-antiformant pair and band pass filtered signal of hypernasal speech vowels contain F_1 only. Because of the presence of extra nasal formant-antiformant pairs, the hypernasal vowels have poor correlation and hence the value of parameter r is less for the hypernasal vowels compared to normal vowels. The final classification decision was based on likelihood ratio test. The method gives average correct identification rates of 94.7% and 94.7% for the normal and hypernasal front vowel /i/, respectively. The average correct identification rates obtained for normal and hypernasal mid vowel /A/ are 93.0% and 93.3% respectively. In another study, Christopher [80] attempted the hypernasality detection by first decomposing the speech signal into intrinsic modes using Empirical Mode Decomposition (EMD) technique and then analyzing the energy in each mode using TEO. Experimental studies on American CLP Craniofacial database showed that the EMD energy based approach yields clear discrimination for normal and hypernasal speech.

Lee et al. [37,81] developed a quantitative index called voice low tone to high tone ratio (VLHR) as a cue for hypernasality detection. The sustained /[a:]/ vowel collected from the eight subjects having hypernasality in their speech was used for the study. The average vowel spectrum was divided into two regions: low-frequency region and high-frequency region by a specific cutoff frequency at 600 Hz. The VLHR index depends on the power of both the frequency regions and it is expressed in decibels (dB) as

$$\text{VLHR} = 10 \times \log_{10}\left(\frac{\text{LFP}}{\text{HFP}}\right), \tag{2.3}$$

where the LFP is the summation of the power from 65 to F_c Hz and the HFP is the summation of the power from F_c to 8000 Hz. The correlation between VLHR feature and the nasalance score obtained from Nasometer device ($r = 0.76, p < 0.01$) was found to be statistically significant. The correlation between the VLHR feature and perceptual score ($r = 0.80, p < 0.01$) was also found to be statistically significant.

Orozco-Arroyava et al. [36] attempted the automatic detection of hypernasality in CP using a set of four different nonlinear dynamic features (NLD) and six entropy measurements computed from speech. The nonlinear dynamic features are Correlation dimension (Dc), Largest Lyapunov Exponents (λ_1), Lempel-Ziv Complexity, Hurst Exponents (H). The entropy features are: Approximate entropy (A_E), Gaussian kernel approximate entropy (GA_E), Simple entropy (S_E), Gauss kernel simple entropy (GS_E), Recurrent period density entropy, Detrended fluctuation analysis. The database used for hypernasality detection consists of five Spanish vowels /a/, /e/, /i/, /o/ and /u/ collected from 238 children in the age range of 5 to 15 years old. The speech samples of 108 children are labeled as normal, whereas the remaining speech samples from 130 children are labeled as hypernasal. The classification between normal and hypernasal speech was done using the SVM classifier. The result showed that the accuracy of the system increases when the nonlinear and entropy features were combined. A best result of 90.56% for /a/ vowel was reported.

The temporal features based on the TEO profile and VLHR capture the centralized energy in the low-frequency region of hypernasal speech. The NLD feature captures the evidence of nonlinear behaviour in the speech production of hypernasal speech and entropy features measure the randomness in hypernasal speech. The temporal features perform good for hypernasality detection, but these features have some limitations. The feature based on the TEO profile requires a cutoff frequency for the low pass filtering of speech, estimation of F_1 location for band pass filtering and a threshold value of likelihood ratio score for the final decision. The cutoff frequency and threshold value depend upon the vowels and the speaker. The change in their values affects the detection accuracy. Also, finding an appropriate cutoff frequency is difficult when the F_1 and F_2 are close as in case of back vowel /u/. So the vowels /u/ was not considered for the hypernasality detection using TEO based feature. The decomposition of speech signal into its intrinsic modes using the EMD method for computing TEO based feature also has the limitation. The EMD method decomposes the speech signal from high frequency to low frequency and it can not separate closely spaced frequencies [82]. So it may not be able to resolve the closely spaced nasal and oral formant frequency component in the low frequency of the hypernasal signal. The VLHR index is proposed only for the vowel /a/ and the value of cutoff frequency 600 Hz is chosen based on the experiment. The value of cutoff frequency may change for other databases. The NLD and entropy features are generally used for the pathological voice disorders which are the problems at the level of the larynx, whereas the hypernasality is considered

as a resonance disorder at the level of vocal tract.

2.4.2 Hypernasality detection using features based on formant analysis

The deviation in hypernasal vowel spectrum due to the addition of nasal formant and antiformant pairs can be captured by the features based on formant strength and frequency locations. For that formant analysis of hypernasal vowels is done to detect the strength and location of nasal and oral formants present in the vowel spectrum. In the literature, various works have been done for hypernasality detection using features based on the formant analysis. Vijayalakshmi et al. [26,83] performed the analysis and detection of hypernasal speech using the feature extracted from the modified group delay (MGD) spectrum. The authors first analyzed the spectral characteristics of nasalized vowels present in /summer/, /sunny/, and /singing/ words and observed that the nasal formants get added in these vowels at various frequency locations. Among the various nasal formants, the formant in the low-frequency around 250 Hz was found most consistent. The effect of nasal formant at 250 Hz on the perception of nasality was tested by the perceptual evaluation of nasality in the synthesized vowels having 250 Hz nasal formant. Later, the authors performed the spectral analysis of hypernasal vowels to confirm the presence of 250 Hz nasal formant. To resolve the closely spaced nasal formant and F_1 in hypernasal vowels, the low-frequency region of the MGD spectrum was analyzed. The MGD spectrum can be computed from the speech signal using the formula,

$$\tau_m(\omega) = \frac{\tau(\omega)}{|\tau(\omega)|}(|\tau(\omega)|)^{\alpha} \tag{2.4}$$

where,

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S(\omega)^{2\gamma}}, \tag{2.5}$$

X_R and Y_R are real part of FFT of $x(n)$ and $nx(n)$ respectively, X_I and Y_I are imaginary part of FFT of $x(n)$ and $nx(n)$ respectively and $\alpha = 0.6$ and $\gamma = 0.9$.

The MGD spectrum has higher frequency resolution compared to the conventional LP or cepstrum spectrum due to the additive property of the group delay function [84]. Based on the low-frequency analysis, the group delay function-based acoustic measure (GDAM) was proposed as a feature for hypernasality detection. The measure was defined as the ratio of absolute value of group delay function at F_1 to the absolute value of group delay function at F_2, where ($F_1 < F_2$) were two most dominant peak frequency in band-limited group delay function (< 800 Hz). The final classification

between normal and hypernasal speech was based on a fixed threshold. If GDAM value was higher than the threshold value, speech was classified as hypernasal speech. If GDAM value was lower than the threshold value, speech was classified as normal speech. The database used for the hypernasality detection experiment consisted of phonemes /a/, /i/ and /u/ uttered by 33 hypernasal speakers and 30 normal speaker and a detection performance of 100% for vowel /a/, 88.78% for vowel /i/ and 86.66% for vowel /u/ was reported. In another study, Lakshmi and Reddy [85] performed hypernasality detection using the LP-based pole modification technique. In this technique, normal and hypernasal vowel spectrum was modeled by 28 order LP spectrum. The high order spectrum resolves the closely spaced nasal and oral formant, but the peaks in the spectrum may be spurious due to pitch harmonics effect. So the strongest peak in the low-frequency region is defocussed and a new signal was resynthesized. The defined hypernasality measure was the maximum of cross-correlation value between the input and the resynthesized speech signal. The hypernasality detection was performed for the /a/ vowel present in the speech samples collected from 25 hypernasal speakers with unrepaired CP and 25 normal speakers. A detection accuracy of 100% was reported. In another work [86], authors performed hypernasality detection using the three-dimensional feature. The frequency locations of the first two highest peaks in the group delay spectrum and the ratio of the group delay at these frequencies are considered as features. The detection was done for /a/, /i/, and /u/ vowels. For each vowel, an accumulated minimum distance classifier and a maximum likelihood classifier were trained separately, and testing was done using these trained classifiers. The database used for the experiment consists of speech uttered form 45 speakers with CP and 26 normal speakers. A classification accuracy of 85% was reported. Qian et al. [87] attempted hypernasality detection by resolving the nasal and oral formants. The LP coefficient roots extraction method was proposed to detect the additional nasal formant in the low-frequency region of the spectrum. The resolved formant frequency was considered as the feature for hypernasality detection. The experiment was carried on 426 (216 normal and 210 hypernasal) phonemes and an accuracy of 95.2% was reported.

The formant analysis based features involve nasal and oral formants detection. To resolve these formants, the MGD spectrum or LP root solving method has been used. The GDAM feature is extracted from the MGD spectrum. The spectrum resolves the closely spaced nasal and oral formants present in the low-frequency region of the spectrum. However, low-pass filtering of speech is done using a predefined cutoff frequency to detect the low-frequency region of the spectrum. The number

of peaks in the low-frequency region may vary based on the value of cutoff frequency. So the spectrum may have spurious peaks which may not represent a formant. The spectrum may also have spikes which suppress the actual formants. The classification using the GDAM feature was done based on the threshold value which varies with vowels and may also vary for other databases. The deforcing of strongest peak in the pole modification technique may suppress the F_1 in case of /i/ and /u/ vowels. So, the technique may not work for hypernasality detection in case of these vowels. The roots solving method to detect formants may also yield the spurious formants.

2.4.3 Hypernasality detection using cepstral features

The cepstral features are used to model the envelope of the spectrum. The envelope of hypernasal vowel spectrum gets affected by the addition of nasal formant and antiformant. In literature, cepstral coefficients extracted from the DFT magnitude spectrum or LP spectrum have been used for hypernasality detection. Rah et al. [33] proposed a quantitative method of hypernasality detection using the LP model. The proposed method was based on the assumption that the LP spectrum with high model order can approximately model the formants as well as antiformants present in the hypernasal vowels spectrum. The assumption was based on the concept that a zero in the spectrum can be modeled using an infinite number of poles and it can be expressed as,

$$1 - z_0 z^{-1} = \frac{1}{1 + \sum_{k=1}^{\infty} z_0^k z^{-k}}, \tag{2.6}$$

The distance between LPCC corresponding to low and high order LP models was proposed as the feature for hypernasality detection. The low order (8-10) LP coefficients capture only the formants present in the normal and hypernasal vowel spectrum. The high order (34-38) LP coefficients capture the formants in the normal vowel spectrum, whereas it captures the formant as well as antiformants in the hypernasal vowel spectrum. Hence, the distance value was found less for normal vowel compared to the hypernasal vowel. For whole speech database, the correlation of the feature with the nasalance score was found to be 0.58, which is better compared to the correlation 0.44 between TEO based feature with the nasalance score. Although the correlation was better compared to TEO based feature, but its value was low 0.58 due to the reduction in formant strength because of the presence of antiformant. The feature gives a high correlation of 0.84 for the part of database having nasalance value greater than the 35. Akafi et al. [88] proposed a quantitative measure for hypernasality detection based on the distance between the sequences of cepstral coefficients. These sequences were corresponding to Autoregressive

(AR) and Autoregressive Moving Average (ARMA) models used to model the hypernasal speech spectrum. The measure was based on the fact that an AR model is not sufficient to model the vocal tract characteristics of children with CP because antiformants are present in their speech spectrum. Therefore, the authors used the ARMA model for modeling the spectrum which can model the formants as well as antiformants in the spectrum. The distance was found more for hypernasal speech compared to normal. The classification was done based on the threshold value and accuracy of 86.67% was reported.

For the normal and hypernasal speech classification using above-mentioned feature, a threshold value of feature is required which may vary for different databases.

2.4.4 Hypernasality detection using combined features

The combined features are the inter combination of temporal, cepstral and features based on formant analysis. Maier et al. [32] proposed a method for the objective measurement and assessment of phonetic disorders present in CP speech and compared the result with the perceptual rating. The work uses the combination of the MFCC and TEO based features for hypernasality detection. The authors performed two experiments. In the first experiment, an automatic speech recognition (ASR) system was developed for the detection of disorders based on the manually created transliteration tests. In the second experiment, detection is performed in a fully automatic way without manual transliteration. The ASR system was developed with the 58 CP children speech database. The speech samples were perceptually assessed for phonetic disorders such as hypernasality, nasalized consonants, pharyngeal backing, laryngeal replacement, and weakened plosives at the phoneme level. The first experiment reported moderate to a good agreement ($\kappa \approx 0.6$) for the detection of all phonetic disorders. The second experiment yield a correlation of (0.81) on the speaker level with the perceptual rating. Rendón et al. [35] attempted the automatic detection of hypernasality in children with CP on five Spanish vowels. The acoustic and noise-based features, Jitter, Shimmer, MFCC, Harmonic to Noise Ratio (HNR), Cepstral-HNR, Normalized Noise Energy (NNE), Glottal to Noise Excitation Ratio (GNE) were computed from the vowels and a high dimensional representation matrix was formed. The most relevant features for better discrimination between the normal and hypernasal vowels were selected using Principal component analysis (PCA) and Linear correlation. A classification performance between 80% to 90% was reported using the Linear-Bayes classifier. Orozco-Arroyava et al. [34] proposed two types of characteristic features, one based on the acoustic, noise and cepstral

features as used in [35] and other based on nonlinear dynamic features proposed in [36] for hypernasality detection. The database used were /a/, /e/, /i/, /o/ and /u/ vowels recorded from 266 children in the age range of 5 to 15 years. Out of 266 children, 110 were having normal speech and 156 were having hypernasal speech. The authors used PCA and heuristic floating search to find the optimal sub-space of features to obtain better discrimination between normal speech and hypernasal speech. A best accuracy of 93.73% was reported using the SVM classifier. Golabbakhsh et al. [39] attempted hypernasality detection by using jitter, shimmer, MFCC, bionic wavelet transform entropy, and bionic wavelet transform energy features. The used speech database consisted of recorded six different sentences uttered by speaker with normal speech and speakers with CP. The SVM classifier was used for the classification between normal and hypernasal speech. The different combination of features was used for the classification. The best performance was reported for the combined MFCC and bionic wavelet transform energy feature. The best-reported accuracy was 85%.

The combined features are giving a better result compared to individual features. The MFCC feature, which is very commonly used in speech processing, was used in these works as the cepstral feature. However, in some studies [45, 89] it has been pointed out that the MFCC feature does not smooth out the pitch harmonics effect in the low frequency of spectrum for high pitch children speech. The inadequate smoothing in lower frequencies may affect the capability of MFCC feature to capture the nasality cues present in lower frequencies. The acoustic features such as jitter and shimmer are used to capture the characteristics of vocal folds. The noise-based features such as HNR are used to capture the influence of noise in hypernasal speech. These features are not directly based on nasality cues.

Besides these features, a work of hypernasality detection has also been done using the whole speech spectrogram as a feature. In that work, Xiyue et al. [90] attempted hypernasality detection based on the feature-independent end-to-end algorithm that uses a convolutional neural network (CNN). The input to the CNN was speech spectrograms. The detection was performed for /a/, /i/ and /u/ vowels and the best result was obtained for /i/ vowel with average F1-scores of 0.95 and 0.97 on children and adults speech database respectively. The result of CNN based detection was found better compared to the result obtained from deep neural network (DNN), backward propagation (BP) neural network, and SVM classifiers. The best result was obtained for the convolutional filter size of 1×8.

In summary, the features used in the above-mentioned works for hypernasality detection (temporal,

formant based and cepstral features) classify the normal and hypernasal speech with good accuracy. But the doctors and SLPs are more interested in the severity rating of hypernasal speech because the severity rating correlates with the velopharyngeal gap size in children with CP. However, very few works have been done related to severity rating which is discussed in the next section.

2.5 Severity grading of hypernasal speech

The rating of nasality in increasing order is called the severity rating. Two severity rating scales are used: 4-point scale and the 6-point scale for hypernasality assessment. The 4-point scale is 0-normal, 1-mild, 2-moderate, 3-severe, and the 6-point scale is 0-normal, 1-mild, 2-mild to moderate, 3-moderate, 4-moderate to severe, 5-severe. However, nowadays the 4-point rating is universally accepted for the hypernasality assessment [1]. The instrumental observation shows that the velopharyngeal gap size increases as the severity of nasality increases in hypernasal speech. The simulation studies show the shift in nasal formant and antiformant pairs as well as in the oral formant as the severity of nasality increases. Very few works have been done by researchers for hypernasality severity grading. Kataoka et al. [40] proposed features based on the one-third octave spectra to quantify the perceived hypernasality in children speech by finding the relation between spectral characteristics and perceived nasality. One-third octave spectra of the isolated vowel /i/ were obtained from 32 children with CP and 5 children without CP. Four experienced listeners rated the severity of hypernasality of the 37 speech samples using a 6-point equal-appearing interval scale. When the average 1/3-octave spectra from the hypernasal group and the normal resonance group were compared, spectral characteristics of hypernasality were identified as increased amplitudes between $F1$ and $F2$ and decreased amplitudes in the region of $F2$. Based on the findings of the children speech, 36 speech samples with manipulated spectral characteristics were used to minimize the influences of voice source characteristics on perceived hypernasality. Multiple regression analysis revealed a high correlation (0.84) between the amplitudes of 1/3-octave bands (1 kHz, 1.6 kHz, and 2.5 kHz) and the perceptual ratings. Increased amplitudes of bands between F_1 and F_2 (1 kHz, 1.6 kHz) and decreased amplitude of the band of F_2 (2.5 kHz) were associated with an increasing perceived hypernasality. He et al. [38] attempted a two-stage process for the detection and severity rating of hypernasal speech. The work was based on the cue that the spectral energy in hypernasal speech gets shifted in the low-frequency region due to the addition of nasal formant and antiformant pairs. In the first stage, the feature energy distribution ratio defined

as the ratio between the energy in the low-frequency region to the total energy in the signal was computed for the hypernasality detection. The ratio is given by,

$$R = \frac{E_{fc}}{E_{\frac{fs}{2}}}, \tag{2.7}$$

where, f_c is the cutoff frequency and f_s is the sampling frequency. The detection was done based on a fixed threshold R_h. The speech was classified as normal if $R < R_h$ and hypernasal if $R > R_h$. In the second stage, hypernasality severity detection of hypernasal classified speech was performed. For that, the Gaussian mixture model was trained using mild, moderate and severe hypernasal speech. The feature used for the hypernasality severity detection was the first formant frequency and the total number of formants in the speech. A classification accuracy of 83% was reported. Liu et al. [91] proposed a hypernasality severity grading method based on the low quefrequency liftering. In this method, low liftering of cepstral coefficient below 90 quefrencies was done and the cepstral spectrum was computed. The spectrum was taken as the feature and the BP neural network based on natural computation was used as the classifier. The experimental study yield a classification accuracy above 80% for four grades (normal, mild, moderate and severe) of hypernasality. Vikram et al. [92] attempted the severity grading of hypernasal speech using the database having recording of sentences uttered by children. The work proposed a method that gives a continuous nasality score for the speech just like the nasometer device. In this method, the two extremely opposite classes of speech having minimum and maximum nasality are modeled using Gaussian mixture model and deep neural network. The oral sentences (rich in vowels, stops, and fricatives) collected from normal speakers were considered for the speech having minimum nasality speech, whereas the nasal sentences (rich in nasals and nasalized vowels) collected from moderate-severe hypernasal speakers were considered for the speech having maximum nasality. The posterior probabilities obtained for oral sentences (as a test speech) were considered as hypernasality scores. The obtained nasality scores show a significant correlation ($p < 0.01$) with respect to perceptual ratings of hypernasality. Further, the hypernasality scores were also used for the detection of hypernasality and a best accuracy of 93.10% was reported. Zhang et al. [93] attempted the hypernasality severity grading using SVM classifier based on vocal tract characteristics which were modeled by stabilized weighted linear prediction (SWLP) and extended weighted linear prediction (XLP) methods. The SWLP method imposes the temporal weights on the closed-phase interval of the glottal cycle, so it was considered a more robust approach for modeling the

vocal tract than LP. On the other hand, the XLP method weights each lagged speech signal separately, so it achieves a finer time scale on the spectral envelope than the SWLP method. The experiment was performed on 4640 Mandarin syllables collected from 60 CP subjects and 20 control normal subjects. The result showed that the spectral envelope of normal speech roll-off faster compared to hypernasal speech in higher frequencies. Also, the correlation coefficients between normal and hypernasal speech for SWLP and XLP based method was found smaller than the LP method. The reported classification accuracies for hypernasality grades (normal, mild, moderate and severe) using the SWLP and XLP methods range from 83.86% to 97.47%. Adam et al. [94] compared the 1/3 octave spectra measure and the VLHR with the perceptual rating. The result shows that the 1/3 octave spectra measure differentiate better between normal and hypernasal speech as well as between different severity levels of hypernasality.

Most of the hypernasality severity detection works were done by multi-class classification of speech which gives the confusion matrix representing the percentage of correctly classified samples for different classes. This kind of classification does not give a nasality score for a child speech as obtained from the nasometer device. Also, the interpretation of the result obtained from the classification requires technical knowledge about the training, testing, and classifier.

2.6 Hypernasality detection in clinical environment

As stated in the introduction of this chapter, hypernasality detection in the clinical environment is done perceptually by SLPs and the perceptual rating is supplemented with the instrumental measurements. The perceptual rating is done by listening to the speech stimuli uttered by the CP children. The SLPs who are highly trained and having five or more years of experience are considered for perceptual evaluation. The SLPs are asked to grade the nasality in the speech at the 4-point scale: normal, mild, moderate and severe. The perceptual judgement is considered a gold standard in hypernasality evaluation. However, the abnormalities in pitch, loudness, voice quality and/or articulation, which occurs in conjunction with the hypernasality [23], may affect the perception of nasality in hypernasal speech [7, 24]. Hence the perceptual rating may vary among the SLPs. The perceptual rating from different clinical centers are also difficult to compare due to the use of different speech stimuli and rating scale [1]. To restrict the limitations of perceptual evaluation, several instrumental evaluation techniques are used to supplement the perceptual evaluation of hypernasality. Unlike the percep-

tual rating, the instrumental techniques give objective measure for hypernasality assessment, and the techniques can be divided into two categories: direct and indirect [10] which are as follows.

2.6.1 Direct assessment techniques

The direct assessment techniques are used to observe the insufficient velopharyngeal port movement during the production of hypernasal speech and gives information about velopharyngeal gap size and shape. This information is used by the clinicians 1) to judge hypernasality and its severity, 2) to evaluate improvement before and after treatment, 3) to select an appropriate treatment procedure for the assessment [14]. Hypernasality assessment decision is done on the basis that more the velopharyngeal gap, severe the hypernasality in the speech. The different direct techniques used for the hypernasality assessments with their advantages and limitations are discussed below.

- **Cephalometry:** In the cephalometry assessment, the velopharyngeal structures are captured with the lateral cephalograms taken by X-ray during the utterance of stimuli [11]. The technique gives the information about the relation of the soft tissues of the nasopharynx to the bony landmarks of the face and the cranium. However, the children are explored to the ionization radiation which may be harmful to them. The information obtained from the technique can be hampered by the presence of multiple shadows. Additionally, the technique only gives the static information in the midsagittal plane which reduces the three-dimensional anatomy into a two-dimensional representation [95].

- **Multiview videofluoroscopy:** Multiview videofluoroscopy is another technique like the cephalometry to observe the velopharyngeal structures using the X-ray during the utterance of stimuli but from a different planes of space [12]. In this technique, the abnormalities of the velum and posterior pharyngeal wall along with the height of velopharyngeal closure can be visualized by the lateral view [96]. The frontal view demonstrates the lateral wall movements, whereas the basal view gives the relationship between the velum and the lateral-posterior aspects of the pharyngeal wall [14]. However, like the cephalometry, this technique also has the ionization radiation, the presence of multiple shadows and two-dimensional representation as limitations.

- **Magnetic resonance imaging:** Magnetic resonance imaging is a more recent imaging technique compare to cephalometry and videofluoroscopy [97, 98]. It gives the images of different plane views with high spatial resolution. In this technique of assessment, the mid-sagittal view

gives the information about length, movement, extensibility of the velum and forward movement of the posterior pharyngeal wall during velopharyngeal closure. The coronal view gives the width of the pharynx and the nature of the lateral pharyngeal wall during velopharyngeal closure. Finally, the axial view offers information about the extent of velopharyngeal closure [99]. The technique is repeatable, reproducible, having synchronized audio with the video and free from ionizing radiation [100,101]. However, the technique is costly, the images are taken in the supine position which is not preferred in the speech production area and it may be uncomfortable to the children [25,102].

- **Nasoendoscopy:** During nasoendoscopic assessment, a fiberoptic scope is inserted into the nasofarynx just above the velum to obtain the bird's-eye view of velum movement during the utterance of stimuli [13]. An appropriate stimuli and good cooperation of the child is needed for the proper nasoendoscopic assessment [103]. In this technique the children are not explored to the ionization radiation and the techniques have a strong correlation between the grade of velopharyngeal insufficiency [13,104]. However, it is an invasive technique which can prevent children to cooperate.

- **Computed tomography:** Computed tomograph scans (CT-scans) gives information about the anatomy of the velopharyngeal system in the axial plane during the utterance of stimuli [105]. The CT-scans images determine the level of velopharyngeal closure and quantify the surficial and deep craniofacial structures [106]. However, the children are exposed to ionizing radiation, only static information in two-dimensional space is obtained by this technique.

- **Ultrasound:** Ultrasound is used to observe the characteristics of lateral pharyngeal walls. In this technique, a transducer in combination with the use of an acoustic coupling gel has to be placed against the neck under the ear or behind the ramus of the mandible [107]. The technique is free from the ionizing radiation. However, the technique has the restricted visibility of the velum [95].

2.6.2 Indirect assessment techniques

The indirect assessment techniques capture the information which can be used to infer about the movement of velopharyngeal port. The technique gives quantitative information about the abnormal functioning of the velopharyngeal mechanism which can be used to judge hypernasality and its

severity. The techniques are based on the principle that the improper movement of the velopharyngeal port affects the normal aerodynamic and acoustic characteristics of speech. So in this technique aerodynamic and acoustic measures are used for the hypernasality assessment.

- **Aerodynamic measurements** The change in aerodynamic characteristics during the production of hypernasal speech is due to the increase of nasal air escape which changes the nasal and oral tracts airflow and air pressure [20]. So, in this technique amount of nasal air escape is measured for the hypernasality assessment. More the nasal air escape, more severe the hypernasal speech. The aerodynamic measurements are done by the two techniques which are as follows.

  - **Nasal and oral airflow technique:** The nasal air escape can be observed by the *mirror-fogging test* proposed by the Glatzel where a cold mirror is held under the nose of the children during the utterance of stimuli [16]. The degree of condensation measured at four concentric circles represents the severity of the nasal escape. The test is simple, noninvasive and inexpensive. However, the reliability and validity of the test depend on the temperature, air humidity resistance of nasal airways and tilting errors [108, 109]. Another device used for the measurement of nasal air emission is the *aerophonoscope*. It consists of three airflow sensors, two for the two nostrils and one for the oral airflow to visualize nasal and oral airflow as well as the voice levels simultaneously. The device also has a display to presents the measured data in the form of a graph. However, the device is held under the nose in front of the mouth which can influence speech. The other instruments like *pneumotachograph* [18], warm wire anemometer [19], *Super Nasal Oral Ratiometry System* [110] can also be used for nasal flow measurement.

  - **Pressure flow technique** This technique is the expansion of the nasal airflow technique for the evaluation of velopharyngeal function during the speech production. In this technique, the velopharyngeal orifice area is determined with the rate of nasal airflow and the differential pressure across the velopharyngeal orifice.

$$\text{Orifice area} = \frac{\text{Volume rate of air through the orifice}}{0.65\sqrt{((2 \times (\text{intraoral air pressure-nasal air pressure}))/\text{density of air})}} \tag{2.8}$$

    To collect the requisite data simultaneously, two flexible catheters, one within the mouth and another in the nostril, are used to collect intraoral and nasal air pressure (mm H_2O)

and transmitted to pressure transducers. Furthermore, airflow is measured (ml/s) by a pneumotachograph connected by plastic tubing to the patients other nostril. However, the needed equipment is not often available and the procedures are technically complex and require substantial cooperation.

- **Acoustic measurements** The acoustic characteristics of hypernasal speech deviates from the normal speech due to the leakage of air through the nose. The researchers attempted to measure the acoustic parameters which can discriminate between normal and hypernasal speech. The acoustic measurements are done by accelerometric, nasometry and spectral analysis techniques for hypernasality assessment. The techniques are discussed below.

  - **Accelerometry:** The technique measures the accelerometer signals from the area of the nasal and oral tracts of the CP children during the utterance of stimuli and proposes a measure to represent the nasality in the speech. The Horiis Oral Nasal Coupling (HONC) Index [21] is such a measure evolved by Horii. To compute the index, two accelerometers, one on the external surface of a nose and others on the throat are placed. The index is the ratio of the nasal accelerometer signal amplitude to the laryngeal accelerometer signal amplitude and it is expressed as,

$$\text{HONC} = \frac{A_{rms}(n)}{k \times A_{rms}(v)}, \tag{2.9}$$

    where, $A_{rms}(n)$ is the root-mean-square amplitude of the nasal accelerometer signal, $A_{rms}(v)$ is the root-mean-square amplitude of the vocal accelerometer signal. k is a constant which corresponds to the value of HONC for the sustained phonation of sound /m/. The value of HONC index ranges from 0 to 1, where 0 represents the oral speech signal and 1 represents sustained sound /m/ [21]. The index has a strong correlation with perceptual evaluation. This assessment technique is noninvasive and can be used for the sustained sounds or continuous speech. However, the HONC index is rarely used in clinical or research settings because its preassembled package is not available commercially [10].

    Another Accelerometry measure is the Nasality Oral Ratio Meter (NORAM) [111]. In this measure, the nasal and oral signal duration is measured by placing the accelerometers on

the nose and larynx. The measure is given by,

$$n = \frac{tN}{tL \times 100}, \tag{2.10}$$

where tN is the duration of the nasal signal, tL is the duration of laryngeal or oral signal and n is the percentage of nasality. NORAM can be used before and after therapy for nasality measurement. However, it has a limited application in clinical and research purposes due to the low inter-and intraobserver reliability and the impossibility to distinguish normal resonance from hypernasality [112].

– **Nasometry:** This acoustic measurement technique is based on the relation between nasal and oral acoustic energy. The measure is called The Oral Nasal Acoustic Ratio (TONAR). The *Nasometer* device marketed by Kay Elemetrics is used for the TONAR measurements. The device is used to measure the percentage of nasalance, i.e., the ratio of nasal acoustic energy to the total oral-plus-nasal energy of speech. The device consists of a headset with a baffle plate containing microphones attached to the top and bottom of the plate and a personal computer (PC). The two microphones and the PC is used for the capturing and storage of the nasal and oral speech separately. The oral and nasal signal passed through a filter with a central frequency of 500 Hz and the bandwidth of 300 Hz. The nasalance percentage is given by,

$$\text{Nasalance } \% = 100 \times \frac{\text{Nasal signal energy}}{\text{Nasal signal energy} + \text{Oral signal energy}}, \tag{2.11}$$

The nasometer device is used at the clinical centers. It is noninvasive, convenient, easy to use and interpret. The quantitative nasalance sore obtained from the device has the subtle differences for language [113], gender [114], age [115] and race [116]. The nasalance score is not significantly affected by the loudness, speech rate and type of oral consonants. The score is affected by the vowel type and it is high for the high vowels. However, the device requires a highly sensitive pressure transducer and different cutoff score to compute sensitivity and specificity. Another important drawback of nasometer is that the instrument cannot be used for the recorded speech.

Because of the above-mentioned limitations of direct and indirect method for hypernasality detection researchers are trying to develop another indirect method based on spectral analysis of speech.

2.7 Summary

This chapter of the thesis reviewed the temporal, formant based and cepstral features proposed in the literature for hypernasality detection and severity grading. The features are based on the nasality cues present in the spectrum of hypernasal speech. The vowels in the hypernasal speech get nasalized due to the coupling of nasal tract with the oral tract. The nasalized vowel spectrum contains the nasal formant and antiformant pairs in the spectrum along with the oral formant. The effect of the presence of nasal formant and antiformant pairs are reported as the nasality cues for hypernasal speech. The reduction in amplitude of formants, especially F_1, shift in the location of formants, presence of the additional peaks in the spectrum and flattening of the spectrum are some important cues that have been proposed in the literature for nasalized vowels. Based on these cues researchers have proposed features for hypernasality detection and severity grading. The temporal features are extracted from the speech in the time domain. Temporal features are capturing the effect of the presence of nasal formant and antiformant pairs in the spectrum on the distribution of energy across the spectrum. The energy gets centralized in the lower frequencies of the spectrum. So, TEO profile based features, VLHR, NLD features and entropy-based temporal features are used to capture the centralized energy in hypernasal vowels. The closely spaced nasal and oral formants are resolved by either high-resolution MGD spectrum or by using the LP spectrum and features based on their strength are proposed for hypernasality detection. The cepstral features such as MFCC and LPCC are used to model the envelope of the hypernasal vowels which get affected due to the presence of nasal formant and antiformant. Further, a combination of temporal, formant based and cepstral features are also used for hypernasality detection. The chapter also reviewed the hypernasality severity detection works attempted in the literature using these features. The severity detection is attempted using the multi-class classification of hypernasal speech.

The chapter also points out the advantages and limitations of the existing temporal, formant-based, and cepstral features for hypernasality detection. The temporal features have the limitation that the features require the low pass filtering of the speech with a predefined cutoff frequency above the F_1. The value of cutoff frequency is different for the different vowels, and for a particular vowel also, the variation in cutoff frequency affects the detection accuracy. Also, the temporal features are only extracted from the speech signal and the residual signal is never explored for hypernasality detection. The formant analysis based features have the limitations that a high-resolution spectrum

is needed to resolve the nasal and oral format and there may be spurious peaks in the spectrum which do not represent the formant. Also, finding the exact location and strength of formants in high pitch children's speech is not considered an easy job. The issues with the explored cepstral features such as MFCC or LPCC are that these features are explored from the DFT magnitude spectrum and phase spectrum is ignored. Further, the DFT magnitude spectrum may do not have enough resolution to resolve the closely spaced nasal and oral tract formant in hypernasal speech. The magnitude spectrum also has the pitch harmonics effect which causes the high variance for the higher coefficients of the MFCC feature. The high variance affects hypernasality detection accuracy. The limitation with the hypernasality severity detection based on multi-class classification is that classification does not give a continuous nasality score just like the nasometer device. Keeping these limitations of the existing features for hypernasality detection, the subsequent chapters of this thesis explores novel features for hypernasality detection and proposes a system for obtaining hypernasality severity score corresponding to children's speech.

3

Hypernasality detection using temporal features

Publications

- **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Detection and assessment of hypernasality in repaired cleft palate speech using vocal tract and residual features," *J. Acoust. Soc. Am.* 146(6), 4211-4223(2019).

Contents

Overview

The hypernasal vowels are characterized by the presence of nasal formant and antiformant pairs in their spectrum which affects the low-frequency characteristics of vowels and temporal characteristics of corresponding residual signals. The strength of harmonics in lower frequencies gets enhanced in hypernasal vowels and the addition of the undesirable signal components around glottal closure instants (GCIs) happens in the residual signal of hypernasal vowels. So in this chapter, vocal tract constriction (VTC) feature which captures the prominence of low-frequencies in hypernasal vowel signal and peak to side-lobe ratio (PSR) feature which captures the residual signal characteristics around GCIs are used for hypernasality detection. The VTC feature is obtained by comparing the speech signal and corresponding zero frequency filtered signal (ZFFS), whereas the PSR feature is obtained by computing the strength of maximum peak and side-lobes in the vicinity of GCIs in residual signal. The hypernasality detection is performed using the combined (VTC+PSR) feature, and its performance is compared with the performance of baseline features. The support vector machine (SVM) classifier is used for hypernasality detection.

3.1 Introduction

The vowels present in hypernasal speech get nasalized due to the coupling of nasal tract with the oral tract. Hence, the spectral analysis of vowels is done to compute the spectral cues denoting nasalization for hypernasality detection. The most important cue for a nasalized vowel is the presence of nasal formant and antiformant pairs in the spectrum, especially in the vicinity of first formant (F_1). The presence of nasal formant around F_1 enhances the strength of low-frequency harmonics in the spectrum and the presence of nasal antiformant diminishes the strength of F_1. Hence the energy distribution in hypernasal speech gets centralized in the low-frequency band. Fig. 3.1 shows the comparison of spectrograms of normal and hypernasal vowels. Fig. 3.1 (a)-(b) show the speech waveforms of normal and hypernasal /a/ vowel, respectively, and (c)-(d) show their corresponding spectrograms. Fig. 3.1 (e)-(f) show the speech waveforms of normal and hypernasal /i/ vowel, respectively, and (g)-(h) shows their corresponding spectrograms. Fig. 3.1 (i)-(j) show the speech waveforms of normal and hypernasal /u/ vowel, respectively, and (k)-(l) show their corresponding spectrograms. From Fig. 3.1 it can be observed that the energy in normal vowels is mainly concentrated around the formants present in the spectrum, but it is mainly present in low-frequency below 500 Hz in hypernasal vowels.

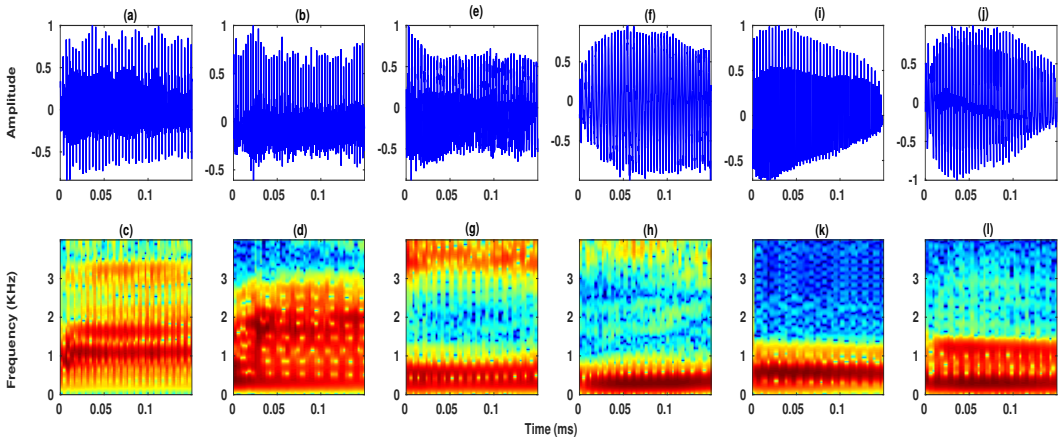

Figure 3.1: Illustration of waveforms of speech signals and the corresponding spectrograms. (a)-(b) are the speech waveforms of a normal and a hypernasal utterance of vowel /a/, respectively. (c)-(d) are their corresponding spectrograms. (e)-(f) are the speech waveforms of a normal and a hypernasal utterance of vowel /i/, respectively. (g)-(h) are their corresponding spectrograms. (i)-(j) are the speech waveforms of a normal and a hypernasal utterance of vowel /u/, respectively. (k)-(l) are their corresponding spectrograms.

The centralization of energy in the low-frequency band deviates the low-frequency characteristics of hypernasal speech compared to normal. Most of the hypernasality detection works such as works based on the TEO profile [31], group delay spectrum [26], VLHR feature [37], and energy distribution ratio (R) feature [38] involve low-pass filtering of the speech with a predefined cutoff frequency to capture the low-frequency characteristics. These features perform well with detection accuracies above 85%. But the problem with these features is that the value of cutoff frequency is different for the different vowels, and for a particular vowel also, the variation in cutoff frequency affects the detection accuracy. Further, it can be noticed that the above-mentioned works are capturing the spectral deviations due to the presence of nasal formants and antiformants for hypernasality detection. But a previous study in [42] shows that the presence of nasal formants and antiformants also deviates the temporal characteristics of linear prediction (LP) residual signal by adding the undesirable signal components in the vicinity of GCIs. However, temporal deviation in the residual signal has not been explored before for hypernasality detection. Motivated from the aforementioned limitations, features are explored in this work for hypernasality detection which (i) do not require predefined cutoff frequency and (ii) can capture the centralized energy in lower-frequencies of the speech signal, and the deviation in temporal characteristics of the residual signal.

In this work, vocal tract constriction (VTC) and peak to side-lobe ratio (PSR) features are used for hypernasality detection. The VTC feature is obtained by comparing the speech signal and the

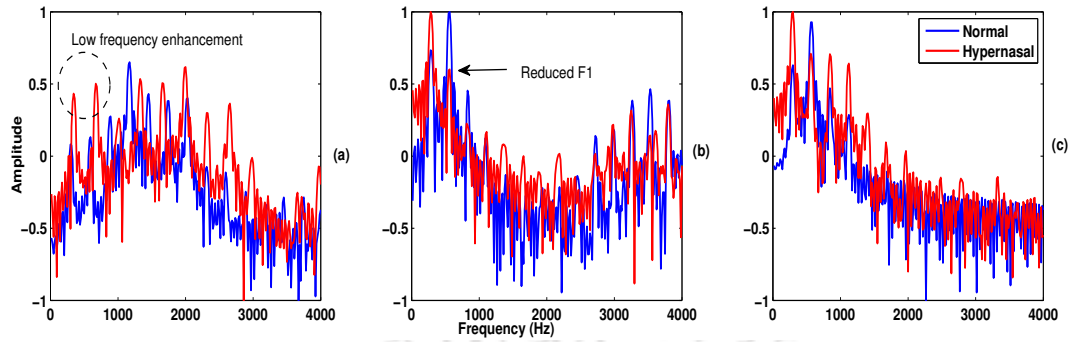

Figure 3.2: Log magnitude spectrum of normal and hypernasal vowels overlapped with each other. (a) for /a/ vowel, (b) for /i/ vowel, and (c) for /u/ vowel. The spectral deviation in the form of dominant low-frequency harmonics below F_1 due to nasal formant and reduction in F_1 strength due to nasal antiformant are illustrated in this figure.

corresponding zero frequency filtered signal (ZFFS). The VTC feature captures the prominence of lower-frequencies in speech signal without involving low-pass filtering with a predefined cutoff frequency. The PSR feature is obtained by computing the strength of maximum peak and side-lobes in the vicinity of each GCIs, also called epoch locations, in LP residual signal. The PSR feature captures the residual signal characteristics around epoch locations. The hypothesis is that the presence of nasal formant and antiformant pairs in the hypernasal speech spectrum deviates the low-frequency characteristics by enhancing the magnitude of low-frequency harmonics and the temporal characteristics of residual signal by adding the undesirable signal components around epochs locations. Hence, the values of VTC and PSR features may differ for hypernasal speech compared to normal speech. Also, as the VTC and PSR features are extracted from the speech and its residual signals in time domain respectively, so these features are categorized as temporal features. These two temporal features may capture the different aspects of nasality present in the hypernasal speech. Hence the hypernasality detection is performed using the combined (VTC+PSR) feature, and its performance is compared with the performance of baseline features. The support vector machine (SVM) classifier is used for hypernasality detection.

The rest of this chapter is organized as follows: Section 3.2 describes the cleft palate database. Section 3.3 discusses the spectral and residual analysis in hypernasal speech. Section 3.4 describes the feature extraction. Section 3.5 describes the experimental results and Section 3.6 summarizes the work and provides the conclusion of the chapter.

3.2 Database

The speech database was collected from All Indian Institute of Speech and Hearing (AIISH), Mysore, India [117]. The database consists of control normal and hypernasal speech. The control normal (CN) speech was collected from 30 children (numbered from 1 to 30) with normal speech and language characteristics. The hypernasal speech was collected from 30 children (numbered from 1 to 30) with repaired CP. There were 12 girls and 18 boys in each group. The children lie in the age group of 7-12 years. Table 3.1 gives the child number, age, and gender of control normal as well as children with repaired CP. Table 3.1 shows that the distribution of age and gender is balanced across CN and CP groups. Each child with CP had adequate language abilities, and none of them had any history of hearing impairment or other congenital syndromes or developmental difficulties. The stimuli considered for recording were the words /papa/, /pipi/ and /pupu/ consisting of vowels /a/, /i/ and /u/ in context of pressure consonant /p/. The speech-language pathologists SLPs from AIISH design the stimuli as per the suggestions given in [1]. Speech was recorded in the Kannada language, which is a Dravidian language spoken in the southern part of India. Each stimulus word was recorded ten times from each child in different sessions which means a total of 300 normal and 300 hypernasal /papa/, /pipi/ and /pupu/ words were recorded. During recording the instructor first uttered each word and then the child repeated it. The recording was conducted in a sound-treated room using Bruel & Kjaer sound-level meter microphone (type 2250-s hand-held analyzer) at sampling frequency 44.1 kHz, 16 bits per sample in .WAV format (Microsoft RIFF file format). Each child was given a comfortable seat in the recording room and the microphone was placed at a distance of 15 cm from the child. Two perceptual tests was done using the recorded sounds. The first test was done to find the degree of hypernasality in each child's speech. For that, all ten recorded /papa/, /pipi/ and /pupu/ words of each child is separately given to three SLPs from AIISH to rate the degree of hypernasality at 4-point scale. The median of three ratings from SLPs for a particular child is considered the childs degree of hypernasality. The severity rating agreement between each pair of SLP is compared using Cohens kappa and Spearmans rank correlation coefficient which is shown in Table 3.2. The correlation values show a good agreement between the SLPs. The SLPs rated 30 Out of 30 CN children with none or normal hypernasality, and 15 children with mild hypernasality, 10 with moderate hypernasal and 5 with severe hypernasal out of 30 repaired CP children. Table 3.1 shows the degree of hypernasality in all CN and CP children's speech.

Table 3.1: Description about the age, gender and degree of hypernasality of control normal as well as the children with repaired CP present in the database.

Control normal children				Children with repaired CP			
Child no.	Age	Gender	Degree of hypernasality	Child no.	Age	Gender	Degree of hypernasality
H1	7	M	Mild	N1	7	M	Normal
H2	7	F	Mild	N2	7	M	Normal
H3	7	M	Mild	N3	7	M	Normal
H4	7	M	Mild	N4	7	F	Normal
H5	7	M	Mild	N5	7	F	Normal
H6	7	F	Moderate	N6	7	M	Normal
H7	8	M	Mild	N7	8	M	Normal
H8	8	M	Moderate	N8	8	M	Normal
H9	8	M	Moderate	N9	8	F	Normal
H10	8	F	Mild	N10	8	F	Normal
H11	8	F	Moderate	N11	8	M	Normal
H12	9	M	Severe	N12	9	F	Normal
H13	9	M	Moderate	N13	9	M	Normal
H14	9	M	Mild	N14	9	F	Normal
H15	9	F	Mild	N15	9	F	Normal
H16	10	M	Severe	N16	10	M	Normal
H17	10	F	Mild	N17	10	M	Normal
H18	10	M	Moderate	N18	10	M	Normal
H19	10	M	Moderate	N19	10	F	Normal
H20	10	F	Severe	N20	10	F	Normal
H21	10	F	Moderate	N21	10	M	Normal
H22	11	M	Mild	N22	11	M	Normal
H23	11	M	Mild	N23	11	M	Normal
H24	11	F	Mild	N24	11	F	Normal
H25	11	F	Mild	N25	11	F	Normal
H26	12	M	Mild	N26	12	M	Normal
H27	12	M	Moderate	N27	12	M	Normal
H28	12	M	Severe	N28	12	M	Normal
H29	12	F	Severe	N29	12	F	Normal
H30	12	F	Moderate	N30	12	M	Normal

The second test was done to find, out of 300 normal and 300 hypernasal recorded /papa/, /pipi/ and /pupu/ words, how many words are having the same decisions by all the three SLPs for normal vs. hypernasal speech rating. For this test, all recorded sounds were randomized in order and given to three SLPs separately for normal vs. hypernasal speech perceptual evaluation of each sound. The speech recordings having the same decisions (full agreement) by all three SLPs are considered for the final database. Table 3.3 gives the total number of each stimulus recorded and the numbers of

Table 3.2: Intra-rater reliability estimation.

Pair of raters	Cohen's kappa	Correlation coefficient
1 and 2	0.76	0.79
2 and 3	0.68	0.70
1 and 3	0.75	0.77

Table 3.3: Description of database in terms of recorded stimuli, its number for normal and hypernasal speech and the number of stimuli having same perceptual decision by three SLPs.

Stimulus	No. of each normal and hypernasal stimuli recorded	No. of stimuli having same decision by three SLPs		Phoneme extracted from stimulus	No. of phonemes having same decision by three SLPs	
		Normal	Hypernasal		Normal	Hypernasal
/papa/	300	271	232	/a/	542	464
/pipi/	300	258	226	/i/	516	452
/pupu/	300	262	242	/u/	524	484

speech recordings having the same perceptual decision by three SLPs per child. From Table 3.3 it can be observed that total 271 normal, 232 hypernasal /papa/ words, 258 normal, 226 hypernasal /pipi/ words and 262 normal, 242 hypernasal /pupu/ words are having the same perceptual decision by 3 SLPs. The manual annotation of vowels /a/, /i/, and /u/ from the words /papa/, /pipi/, and /pupu/, respectively was done by the SLPs. The annotation is done by the careful visualization of speech waveform and spectrogram using the Wavesurfer tool [118]. Fig. 3.3 shows the manually annotation of vowel /a/ in the word /papa/. It can be observed that the region of vowels /a/ was considered from the label mark /p/ to label mark /a/. Table 3.3 shows the number of /a/, /i/, and /u/ phonemes present in the database. As a stimuli have two vowel phoneme, so the database consist of a total of 542 normal, 464 CP phoneme /a/, 516 normal, 452 CP phoneme /i/, and 524 normal, 484 CP phoneme /u/ with the same decisions by all the three SLPs.

3.3 Analysis of hypernasal speech

In this chapter, hypernasality detection is done using VTC and PSR features. The VTC feature capture the deviation in low-frequency characteristics of hypernasal speech and the PSR feature captures the deviation in the LP residual signal of the hypernasal speech. Hence in this section, the spectral and residual analysis of hypernasal speech is done to highlight the deviations.

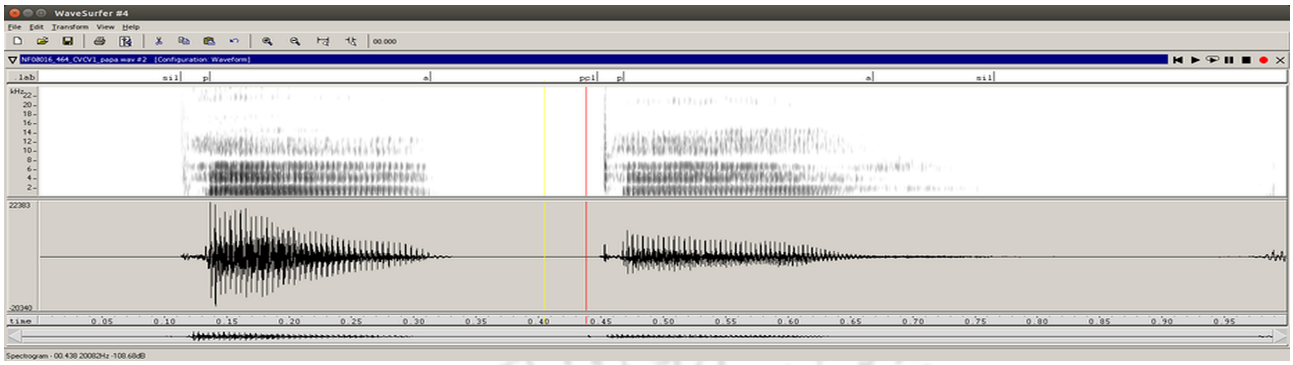

Figure 3.3: Illustration of manually annotation of vowel /a/ in the word /papa/ using Wavesurfer tool.

3.3.1 Spectral analysis

The abnormal coupling of the nasal and oral tracts during the production of hypernasal speech introduces nasal formant and antiformant pairs in the vowel spectrum. This happens at the natural frequencies of the coupled nasal tract and the sinuses attached to it. The natural frequencies of the nasal tract lie in the ranges of 450 to 650 Hz and 1800 to 2400 Hz [119], whereas the natural frequencies of the sinuses lie in the range of around 400 Hz and 1300 Hz [55]. The nasal formants around 400 Hz and in the range of 450 to 650 Hz are responsible for the prominence of low-frequency harmonics in hypernasal vowels compared to normal. The addition of antiformant between nasal formant and F_1 is responsible for the reduction of F_1 magnitude in hypernasal vowels [29]. Fig. 3.2 (a)-(c) show the overlapped DFT spectrum of normal and hypernasal /a/, /i/, and /u/ vowels, respectively to illustrate the low-frequency prominence and reduction of F_1 strength in hypernasal vowels compared to normal. From Fig. 3.2, it can be observed that the low-frequency harmonics below F_1 are prominent and the magnitude of F_1, which lies around 1000 Hz for low vowel /a/ and in the range of 400-600 Hz for high vowels /i/ and /u/, is low. To support the prominent lower-frequencies point, the low-frequency band below 600 Hz energy relative to the total energy is measured for all normal and hypernasal vowels in the database. The measure gives centralized energy at lower frequencies. Fig. 3.4 shows the box plots of centralized low-frequency energy in normal and hypernasal vowels present in the database. Fig. 3.4 (a)-(c) are plotted for /a/, /i/, and /u/ vowels respectively. The box plots indicate that the median of centralized energy is high for hypernasal vowels compared to normal. The Kolmogorov-Smirnov (KS) test for normality shows that the centralized energy distribution is Gaussian ($p < 0.01$) for normal and hypernasal /a/, /i/, /u/ vowels. Hence, one-way analysis of variance (ANOVA) test is

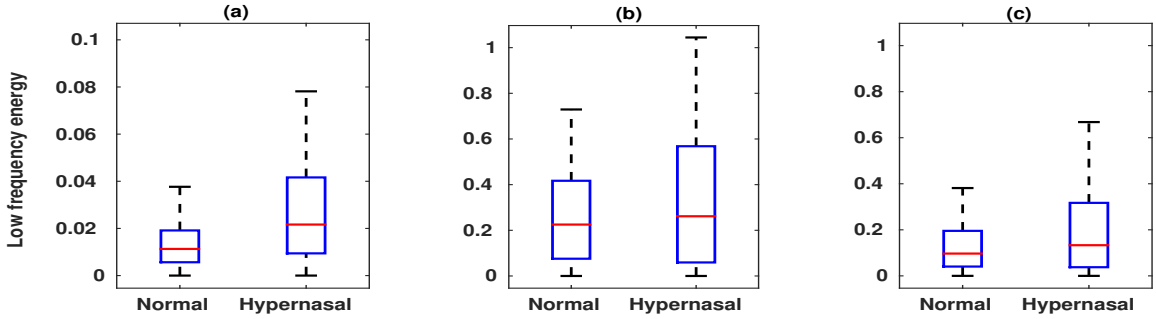

Figure 3.4: Box plots showing the centralized low-frequency energy in normal and hypernasal realizations of vowels (a) /a/, (b) /i/, and (c) /u/.

Table 3.4: Mean, standard deviation (std) and ANOVA test result (F-value and p-value) of centralized energy values corresponding to all normal and hypernasal vowels in the database.

Vowel	Normal		Hypernasal		F-value	p-value
	mean	std	mean	std		
/a/	0.04	0.09	0.08	0.12	831.36	<0.001
/i/	0.32	0.22	0.45	0.31	988.38	<0.001
/u/	0.16	0.14	0.29	0.26	1595.90	<0.001

performed to evaluate the difference between the means corresponding to centralized energy in normal vowels and centralized energy in hypernasal vowels. Table 3.4 shows the mean, standard deviation (std) and the ANOVA test result (F-value and p-value) of centralized energy values corresponding to all normal and hypernasal vowels in the database. The high F-value and a significant p-value ($p < 0.001$) shows that the low-frequency band centralized energy in normal and hypernasal vowels are significantly different. Further, the high mean value of low-frequency band centralized energy for hypernasal vowels compared to normal suggests that lower frequencies are more prominent in hypernasal vowels compared to normal. The spectral analysis thus confirms that there is a significant spectral deviation in hypernasal vowels compared to normal vowels in the form of prominent lower frequencies.

3.3.2 Linear prediction residual analysis

The LP is an all-pole model [120]. The LP spectrum is widely used for the estimation of formants in the vowels spectrum. However, the spectrum is unable to estimate the formants and its bandwidth correctly in the hypernasal vowel spectrum due to closely spaced nasal and oral formants and the

presence of antiformants. A previous study shows that the inaccurate estimation of formants and their bandwidths due to the presence of nasal formants and antiformants add undesirable signal components in the LP residual signal around the epoch locations [42]. The presence of undesirable signals in the LP residual signal can be explained by following mathematical expressions [42].

Consider the vocal tract as a single resonator having impulse response given by

$$v(nT) = \begin{cases} 0; & n < 0 \\ 1; & n = 0 \\ -\hat{a_1}v(nT-T) - \hat{a_2}v(nT-2T); & n > 0 \end{cases} \tag{3.1}$$

where, n and T are the sampling instant and sampling interval, respectively. The reciprocal of the z-transform of $v(nT)$ is given by

$$V^{-1}(z) = 1 + \hat{a_1}z^{-1} + \hat{a_2}z^{-2} \tag{3.2}$$

where, $\hat{a_1}, \hat{a_2}$ are the LP coefficients. Let the new estimated LP coefficients for $v(nT)$ signal after considering the presence of nasal formant and antiformant pairs in hypernasal speech spectrum be $a_1 = \hat{a_1} + \tilde{a_1}, a_2 = \hat{a_2} + \tilde{a_2}$, where, $\tilde{a_1}$ and $\tilde{a_2}$ are the error in $\hat{a_1}$ and $\hat{a_2}$, respectively. Then, the modified inverse filter transfer function $A(z)$ will be given by

$$A(z) = 1 + \hat{a_1}z^{-1} + \hat{a_2}z^{-2} + \tilde{a_1}z^{-1} + \tilde{a_2}z^{-2} \tag{3.3}$$

and the modified z-transform of LP residual $e(nT)$ for the signal $v(nT)$ can be written as

$$E(z) = A(z)V(z) = 1 + \tilde{a_1}z^{-1}V(z) + \tilde{a_2}z^{-2}V(z) \tag{3.4}$$

Hence, the LP residual $e(nT)$ of hypernasal speech can be written as

$$e(nT) = \delta(t) + \tilde{a_1}v(nT-T) + \tilde{a_2}v(nT-2T) \tag{3.5}$$

It can be noticed from equation 3.5 that the LP residual signal of hypernasal speech contains the scaled and delayed versions of the original signal $\tilde{a_1}v(nT-T) + \tilde{a_2}v(nT-2T)$ (which are called undesirable signal components) in addition to the impulse around the epoch locations. The undesirable signal components cause deviation in the temporal characteristics of the residual signal of the hypernasal vowel compared to the normal vowel.

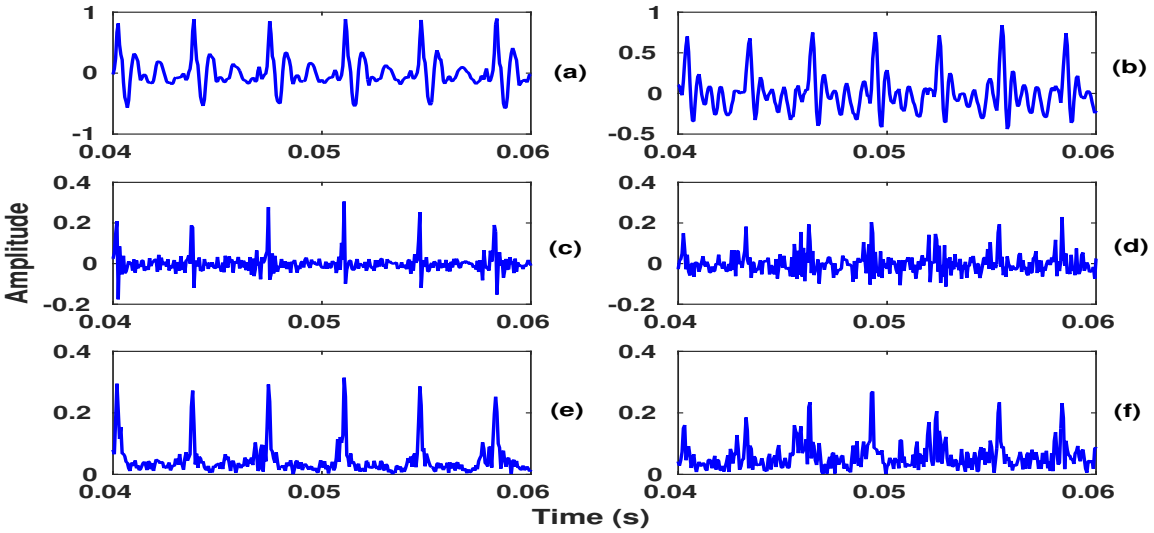

Figure 3.5: Illustration of difference in the nature of residual signals of normal and hypernasal realization of /i/ vowel. (a), (c), and (e) are the speech waveform of normal /i/vowel, LP residual signal, and HE of LP residual signal, respectively. (b), (d), and (f) are the speech waveform of hypernasal /i/ vowel, LP residual signal, and HE of LP residual signal, respectively.

The inverse LP filtering of speech signal gives the residual of the signal [120]. The large amplitude fluctuations, either in positive or negative polarity in the LP residual is the location of epochs. The difficulty due to polarity change can be overcome by using the Hilbert envelope (HE) of LP residual. The HE $h_e(n)$ of LP residual signal $e(n)$ is defined as [121]

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{3.6}$$

where, $e_h(n)$ is the Hilbert transform of the $e(n)$, and is computed as follows:

$$e_h(n) = \text{IDFT}\left\{E_h(k)\right\} \tag{3.7}$$

where,

$$E_h(k) = \begin{cases} -jE(k), & k = 0, 1, ...(\frac{N}{2}) - 1 \\ jE(k), & k = (\frac{N}{2}), (\frac{N}{2}) + 1, ..., (N-1) \end{cases} \tag{3.8}$$

and $E(k)$ is the DFT of the residual signal $e(n)$ and N is the number of points for computing DFT.

Fig. 3.5 illustrates the difference in the residual signal of normal and hypernasal /i/ vowels. Fig. 3.5

(a), (c), and (e) show the speech waveform of normal vowel, LP residual signal, and HE of LP residual signal, respectively. Fig. 3.5 (b), (d), and (f) show the speech waveform of hypernasal vowel, LP residual signal, and HE of LP residual signal, respectively. It can be observed from Fig. 3.5 (d) that the undesirable signal components in the form of additional peaks of both negative and positive polarity around the epoch locations are present in the LP residual of hypernasal speech. These undesirable signal components are responsible for deviation in the residual signal of hypernasal speech compared to residual signal of normal speech. For a better representation of the deviation present in the residual signal, the HE of residual signal is shown in Fig. 3.5 (e) for normal speech and Fig. 3.5 (f) for hypernasal speech. It can be observed from the HE of residual signals that the strength of peaks at epoch locations is lower and the strength of peaks around epoch locations, which are also called side-lobes, are higher for hypernasal vowel compared to normal vowel.

3.4 Feature extraction

3.4.1 Vocal tract constriction feature

The VTC feature measures the prominence of lower-frequencies in speech signal by matching the speech signal and corresponding ZFFS using a cosine kernel [122]. The ZFFS is obtained by first filtering the speech signal through a 0 Hz resonator (also called zero frequency filter) and then removing the trend from the output using a window equal to the average pitch period [123]. In the time domain, the ZFFS is a low-frequency prominent sinusoidal-like signal oscillating with a frequency approximately equal to (f_0). In case of children speech, especially CP children, it is observed that sometimes the ZFFS is not purely sinusoidal. The reason may be high source-filter interaction due to the coupling of nasal tract and high pitch perturbation. So the modified method for finding the ZFFS proposed in [124] is used here. In this method, the differenced speech signal is passed through a cascade of three zero-frequency resonators (ZFRs). The method involves two steps: first, compute the output of a cascade of three ideal digital resonators at 0 Hz.

$$y(n) = -\sum_{k=1}^{6} a_k y(n-k) + x(n) \tag{3.9}$$

where, $a_1 = +6, a_2 = -15, a_3 = +20, a_4 = -15, a_5 = +6, a_6 = -1$ and $x(n)$ is the differenced speech signal. Then, remove the trend from the output i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n) \tag{3.10}$$

where, $\bar{y}(n) = (1/(2N+1)) \sum_{n=-N}^{N} y(n)$ and $2N+1$ correspond to the average pitch period computed over a longer segment of speech. The trend removed signal $\hat{y}(n)$ is the ZFFS.

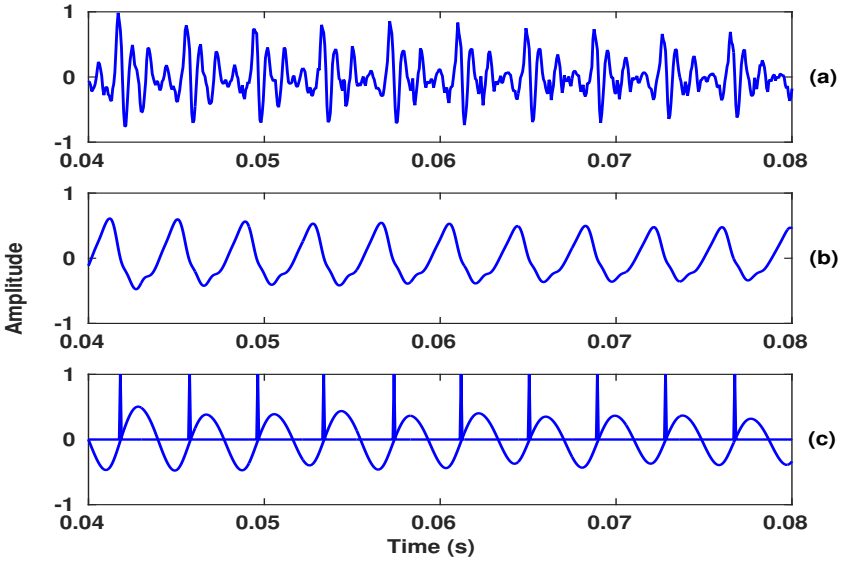

Figure 3.6: Illustration of waveforms of (a) speech signal (b) ZFF signal (c) modified ZFF signal.

The location of epochs is the positive zero crossings of the ZFFS [123]. Fig. 3.6 (a)-(b) show the waveform of hypernasal vowel /i/ and corresponding ZFFS, respectively. It can be observed from Fig. 3.6 (b) that the ZFFS is not purely sinusoidal and it also has a constant delay. The delay introduces the phase difference between the signal and the corresponding ZFFS. Fig. 3.6 (c) shows the modified ZFFS and epoch locations. The modified ZFFS is purely sinusoidal and the delay is adjusted by the constant shifting of ZFFS to make the speech signal and corresponding ZFFS in the same phase. The VTC feature is computed by finding the epoch to epoch correlation between the speech signal and corresponding modified ZFFS using the cosine kernel given by

$$k = \frac{< x'(n), y'(n) >}{||x'(n)||||y'(n)||} \tag{3.11}$$

where, $x'(n)$ and $y'(n)$ are epoch to epoch speech and ZFFS, respectively. The cosine kernel value k is the measure of the match between the two signals. Since the hypernasal speech has the dominant low-frequency components, it will exhibit high similarity with low-frequency dominant sinusoidal like ZFFS. Hence the value of VTC feature k will be high for hypernasal signal compared to the normal

signal.

Fig. 3.7 shows the VTC feature for /i/ vowel. Fig.3.7 (a), (c), and (e) represent the speech waveform for normal vowel, ZFF signal, and VTC feature value, respectively. Fig. 3.7 (b), (d), and (f) represent the speech waveform for hypernasal vowel, ZFF signal, and VTC feature value, respectively. The VTC feature is computed between each successive epoch interval.

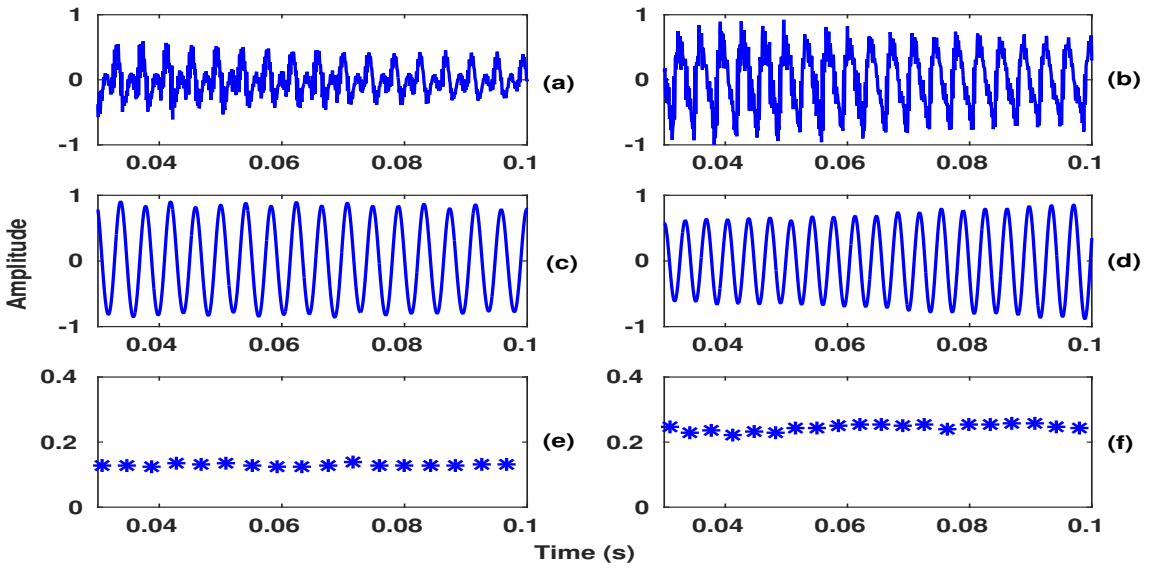

Figure 3.7: VTC feature for normal and hypernasal /i/ vowel. (a), (c), and (e) are the speech waveform for normal /i/ vowel, ZFF signal, and VTC feature value, respectively. (b), (d), and (f) are the speech waveform for hypernasal /i/ vowel, ZFF signal, and VTC feature value, respectively.

3.4.2 Peak to side-lobe ratio

The residual signal is ignored for hypernasality detection because the hypernasality is considered as a system disorder or resonance disorder [3] and the residual signal represents the excitation source characteristics [27]. But as explained in the Subsection 3.3.2, the presence of nasal formant and antiformant pairs in hypernasal vowels spectrum give the deviation in its residual signal, this work extracts PSR feature from the HE of LP residual for hypernasality detection.

To define the PSR feature, the nature of side-lobes in HE of residual signal of normal and hypernasal signal round the epoch location is analyzed. For that, the maximum strength peak locations around each epoch locations in HE of LP residual signals are detected and then a frame of size 3 ms from HE of LP residual signal centered around each peak location is selected. Each frame is normalized by

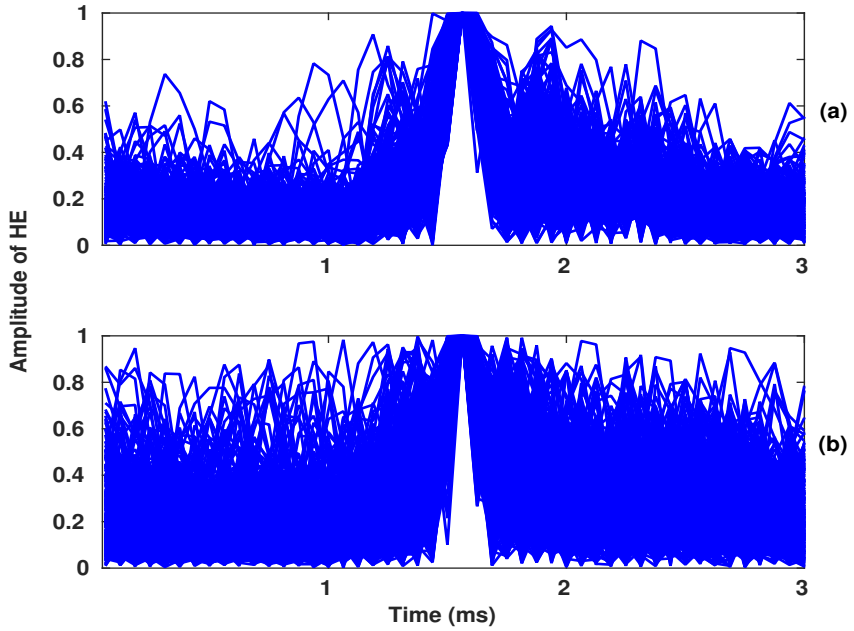

Figure 3.8: 3 ms duration of superimposed segments of HE of LP residual in the vicinity of epoch location. (a) for normal and (b) for hypernasal /i/ vowel.

dividing each sample value in the frame by the maximum value in that frame. The equal number of such frames for normal and hypernasal speech are superimposed and plotted in Fig. 3.8 (a) for normal and (b) for hypernasal /i/ vowel. It can be observed from Fig. 3.8 that the strength of side-lobes in hypernasal vowels is more compared to normal vowels.

PSR feature is defined as $\frac{P}{\mu}$, where, P is the strength of peak at the epoch location and μ is the mean of strength of side-lobes around the peak. The samples from 2 to 3 ms duration in 3 ms window centered at the epoch location are considered as side-lobes [121]. Fig. 3.9 shows the PSR feature for normal and hypernasal /i/ vowel. Fig. 3.9 (a), (c), and (e) are the speech waveform of normal /i/ vowel, HE of LP residual signal and PSR value, respectively. Fig. 3.9 (b), (d), and (f) are the speech waveform of hypernasal /i/ vowel, HE of LP residual signal and PSR value, respectively. It can be observed from Fig. 3.9 (e) and (f) that the value of PSR feature is low for hypernasal vowels compared to normal. The reason for a low value is the decrease in the strength of peaks at epoch location and increase in the strength of side-lobes around the peaks.

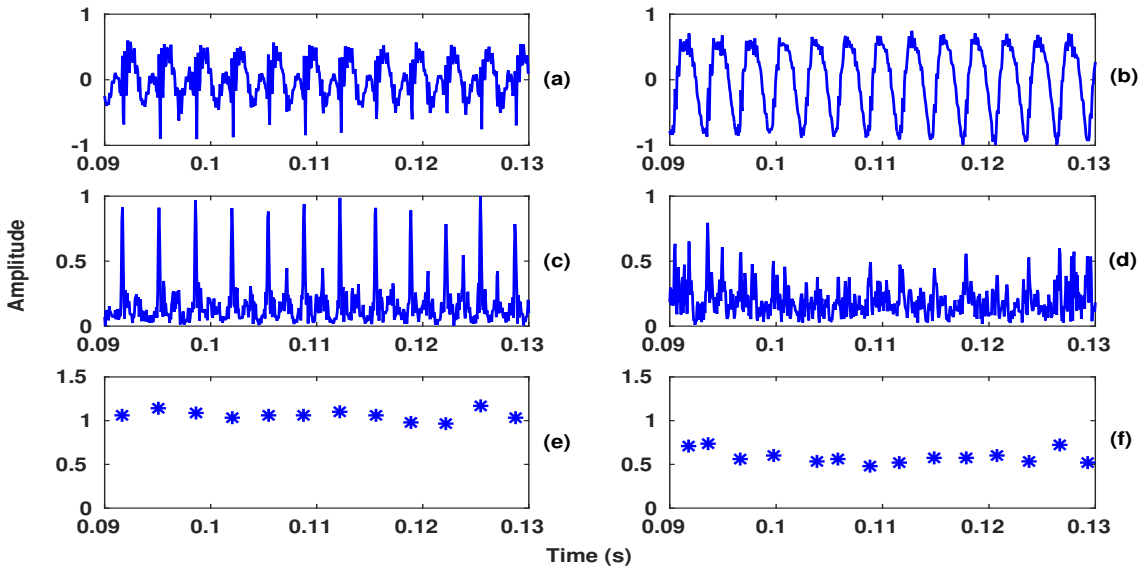

Figure 3.9: PSR feature for normal and hypernasal /i/ vowel. (a), (c), (e) are the speech waveform of normal /i/ vowel, HE of LP residual signal and PSR value respectively. (b), (d), (f) are the speech waveform of hypernasal /i/ vowel /i/, HE of LP residual signal and PSR value respectively.

3.4.3 Statistical analysis of features

Fig. 3.10 (a)-(c) show the box plots of VTC features for normal and hypernasal /a/, /i/, /u/ vowels, respectively. Fig. 3.11 (a)-(c) show the box plots of PSR features for normal and hypernasal /a/, /i/, /u/ vowels, respectively. The features are evaluated from the entire speech database. The box plots show that the median of VTC feature is higher for the hypernasal vowel compared to normal, whereas it is lower for PSR feature. The KS test for normality shows that the distributions of VTC and PSR features for normal and hypernasal /a/, /i/, /u/ vowels are Gaussian ($p < 0.01$). Table 3.5 shows the mean and std of VTC and PSR features along with the result of the ANOVA test for normal and hypernasal /a/, /i/, /u/ vowels. The ANOVA test shows high F-value and significant p-value ($p < 0.001$) which indicates that features for normal and hypernasal speech are differentiable. The clear distinction between the values of VTC and PSR features for normal and hypernasal vowels gives the motivation to use these features for the classification of normal and hypernasal speech.

3.5 Experiments and results

This section of the chapter compares the performance of (VTC+PSR) and baseline features for normal and hypernasal speech classification. The classification results are evaluated using the support

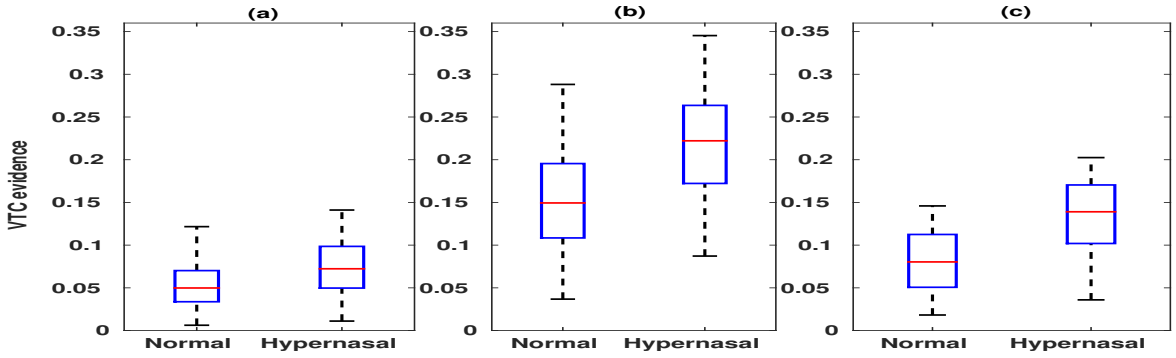

Figure 3.10: VTC feature box plots for normal and hypernasal realization of vowels. (a) /a/ , (b) /i/ , and (c) /u/.

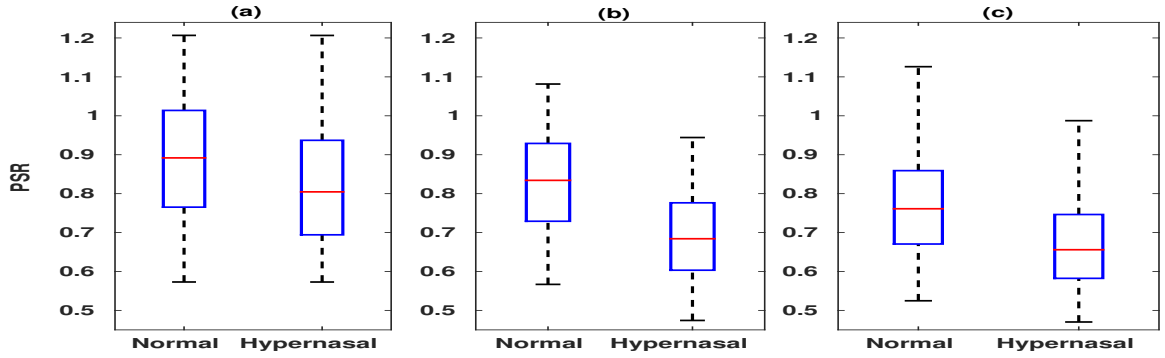

Figure 3.11: PSR feature box plots for normal and hypernasal realization of vowels. (a) (a) /a/, (b) /i/, and (c) /u/.

vector machine (SVM) classifier.

3.5.1 Baseline features

- **Energy distribution ratio feature:** This feature is defined as $R = \frac{E_{fc}}{E_{fs/2}}$, where E_{fc} represents the energy of a speech frame in frequency band of 0 Hz to cutoff frequency fc Hz, and $E_{fs/2}$ represents the total energy of the frame [38]. The cutoff frequency of $fc = 700$ Hz is considered in this work.

- **Group delay function based acoustic measure (GDAM):** This feature is defined as the ratio of absolute value of group delay function at F_1 to the absolute value of group delay function at F_2 where $(F_1 < F_2)$ are two most dominant peak frequency in band-limited group delay function (< 800 Hz) [26]. The steps to compute the group delay function is explained in [26].

Table 3.5: Mean, standard deviation (std) and the result of ANOVA test of VTC and PSR features for normal and hypernasal vowels present in the entire database.

Vowel	Feature	Normal		Hypernasal		F-value	p-value
		mean	std	mean	std		
/a/	VTC	0.06	0.05	0.10	0.06	1429.13	<0.001
	PSR	0.89	0.20	0.81	0.21	716.42	<0.001
/i/	VTC	0.16	0.08	0.22	0.08	1757.54	<0.001
	PSR	0.82	0.16	0.71	0.15	2134.30	<0.001
/u/	VTC	0.11	0.07	0.17	0.09	2601.93	<0.001
	PSR	0.76	0.15	0.67	0.12	1815.24	<0.001

- **TEO based feature (TEOF):** This feature is the frame by frame difference between the TEO profile for low pass and bandpass filtered speech signal [31]. It is quantified in the form of correlation coefficient $r = \frac{C}{\sigma_{LPF} * \sigma_{BPF}}$, where σ_{LPF} is the low-pass filtered profile and σ_{BPF} is the high-pass filtered profile.

- **MFCC feature:** MFCC feature is widely used in automatic speech and speaker recognition. Psychophysical studies have shown that the human ears have better frequency resolution at lower-frequencies of speech spectrum compared to the higher frequencies. Human ears follow the Mel-scale instead of linear scale across the audio spectrum. Mel-scale relates the actual measured frequency with the perceived frequency of a tone. Hence to match the human ear, Mel-scale is incorporated while extracting the MFCC feature from the speech spectrum. A given frequency can be converted into a Mel frequency by using the formula

$$M(f) = 1125 \times \log(1 + \frac{f}{700}) \tag{3.12}$$

  The MFCC feature has been used for the hypernasality detection [32]. This is because the feature models the vocal tract characteristics and has better frequency resolution in lower frequencies where the nasality evidence is mainly present.

  To extract the MFCC feature from the speech signal, the segmentation of signal into frames is done. Each frame is multiplied with the Hamming window, and the power spectrum of resulting windowed signal is computed by taking the DFT. In the next step, Filter-bank energies are obtained from the power spectrum by applying the Mel-frequency filter-bank to the power spectrum. The filter-bank is a set of triangular filters whose cutoff frequencies on the Mel-scale

are calculated using Equation 3.12. The discrete cosine transform (DCT) of the logarithm of FBEs gives the cepstral coefficients [125]. Generally, the lower coefficients are kept as MFCC features.

- **Acoustic (Aco), noise (Noi) and cepstral feature:** The Jitter and Shimmer as acoustic features, Harmonics to Noise Ratio (HNR) [126], Cepstral-HNR (CHNR) [127], Normalized Noise Energy (NNE) [128] and Glottal to Noise Excitation Ratio (GNE) [129] as noise features, and MFCC as cepstral feature has been used in literature for hypernasality detection [34, 36]. Aco and Noi features capture the lack of control of vocal fold, the problem in vocal fold movement and the influence of noise in hypernasal speech, whereas the MFCC feature captures the vocal tract shape during speech production.

- **Non-linear Dynamics (NLD) and entropy feature:** NLD and entropy features have been used together in literature for hypernasality detection [35]. The NLD features used are: Correlation Dymension (D_c) [130], Largest Lyapunov Exponent (λ_1) [131], Lempel-Ziv Complexity (LZ) [132], and Hurst Exponent (H) [133]. Entropy feature used are: Approximate entropy (A_E) [134], Gaussian Kernel Approximate entropy (GA_E), Sample Entropy (S_E) [135], Gaussian Kernel Sample Entropy (GS_E), Recurrent Period Density Entropy ($RPDE$) [136], and a measure derived from a Detrended Fluctuation Analysis (DFA). The NLD features capture evidence of non-linearities in vocal fold vibration and suffering of CP children with the problem of vocal tract articulation and vocal fold movement, whereas the entropy features measure the randomness in hypernasal speech.

3.5.2 Support vector machine classifier

SVM is a kernel machine that has been used for a variety of tasks, including the classification of normal and hypernasal speech [32,35]. SVM is a binary classifier in its basic form that learns a decision boundary by maximizing the margin between two classes. For that, it attempts to maximize the margin between the classes. A kernel function transforms the original input set to a high dimensional feature space where the input samples are linearly separable [137]. If the transformer function is non-linear it allows non-linear decision boundaries between the classes.

Training of a SVM amounts to maximization of the expression,

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}), \tag{3.13}$$

using a set of N training vectors $x_1, x_2, ..., x_N$ with corresponding known targets $y_1, y_2, ..., y_N$ subjected to $\sum_{i=1}^{N} \alpha_j y_i = 0$ and $0 \leq \alpha_i \leq c$, where K is the kernel function, α_i are a set of adjustable weights and b is a bias, both to be learned during training c is a user-defined parameter. The set of $\mathbf{x_i}$ for which $\alpha_i \geq 0$ is called support vectors. If $\mathbf{x_i}$ is a point in input space with unknown classification, then,

$$\hat{y}' = sign\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x_i}, \mathbf{x}') + b\right), \tag{3.14}$$

where $y_i \in \{-1, 1\}$ and $\hat{y}'$ is the predicted class of point $\mathbf{x}'$. The output of the SVM classifier for each speech sample is mapped into probability estimates between 0 to 1 using logistic regression to find the probability that the sample belongs to a specific class [138]. This probability value is called score.

3.5.3 Experiments

The VTC and PSR features are extracted from the speech signal at every epoch location, which is further processed framewise by taking the average of all the values within the frame. The baseline features (R, GDAM, TEOF, MFCC, (Aco+NOi+MFCC), and (NLD+Entropy)) are extracted from the speech signal with the frame size of 20 ms and frameshift of 10 ms. A total of 30 filterbanks are used to compute 30 cepstral coefficients and only lower 13 coefficients ($C_1 - C_{13}$) are kept as a 13-dimensional MFCC feature. The SVM based classification between normal and hypernasal speech is done with radial basis function (RBF) kernel. The LIBSVM toolbox [139] is used in thesis for the SVM based classification. The SVM scores are mapped into probability estimates (0 to 1) using logistic regression. For each vowel, 5 training-testing sets of normal and hypernasal speech are prepared. Each set contains randomly selected 24 normal and 24 hypernasal (12 mild + 8 moderate + 4 severe) children data for training and remaining 6 normal and 6 hypernasal (3 mild + 2 moderate + 1 severe) children data for testing. Hence, each set is speaker-independent, i.e., none of the speaker's speech is used in train and test at the same time. Table 3.6 shows the children in training and testing for each set for both normal and hypernasal speech. The 1 to 30 numbering of children is same as given in Table 3.1. The two-class SVM model is trained and tested for all 5 sets per vowel. All combinations of RBF kernel parameters (c, γ) in the range $c = [2^{-10}, 2^{-8}, ..., 2^{+8}, 2^{+10}]$ and $\gamma = [2^{-10}, 2^{-8}, ..., 2^{+8}, 2^{+10}]$

Table 3.6: Section of children's speech data for five training-testing sets.

Set no	Normal		Hypernasal	
	Children's speech in training	Children's speech in testing	Children's speech in training	Children's speech in testing
1	N2-N8, N10-N12, N14-N21, N23, N25-N28, N30,	N1 , N9, N13, N22, N24, N29	H2-H8, H10-H23, H26 H28, H30	H1, H9, H24-H25, H27, H29
2	N1, N3-N4, N6-N9, N11-N15, N17-N25, N27, N29, N30	N2, N5, N10, N16, N26, N28	H1-H13, H15-H20, H24-H26, H29-H30	H14, H21-H23, H27-H28
3	N1, N4-N9, N11-N17, N19-N22, N24-N25, N27-N30	N2-N3, N10, N18, N23, N26,	H1-H5, H7-H12, H14, H16-H19, H21-H22, H24, H26-H30	H6, H13, H15, H20, H23, H25
4	N3-N10, N13-N26, N29-N30	N1-N2, N11-N12, N27-N28	H1-H13, H17-H18, H20-H24, H26, H28-H30	H14-H16, H19, H25, H27
5	N1-N7, N12-N24, N26-N28, N30	N8-N11, N25, N29	H1-H4, H6-H7, H9, H11, H13-H20, H22, H24-H30	H5, H8, H10, H12, H21, H23

is considered during classification [140]. c is the parameter for the soft margin cost function, which controls the influence of each support vector and γ is the free parameter of the Gaussian radial basis function. The best accuracy obtained in this range of c and γ is reported as a performance measure for that particular set for each vowel.

3.5.4 Results

The results of normal and hypernasal speech classification are presented at two levels: frame level and phoneme level. The parameters accuracy (Acc), sensitivity (Sen), and specificity (Spe) are used to present the result. Sensitivity is the percentage of total hypernasal speech frames/phonemes which are correctly detected, specificity is the total normal speech frames/phonemes which are correctly detected, and accuracy is the percentage of hypernasal and normal speech frames/phonemes which are correctly detected. Phoneme level results are derived from the frame-level results. It is done by using the class labels given by the SVM classifier on a majority basis i.e a phoneme will belong to a particular class if majority of its frames belong to that particular class. The results are shown in Table 3.7, Table 3.8, and Table 3.9 for /a/, /i/ and /u/ vowels, respectively. The left half of each table shows the result at the frame level, and the right half shows the result at the phoneme level. Each table shows the results for VTC feature, PSR feature, combined (VTC+PSR) feature and baseline features. For a particular feature, the average result of all five training-testing sets in terms of mean and std is shown in each table.

From the results for /a/, /i/ and /u/ vowels, it can be observed that the individual performance

Table 3.7: Hypernasality detection using VTC and PSR features for /a/ vowel.

	Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**AUROC**	**Acc (%)**	**Sen (%)**	**Spe (%)**
VTC	70.00±0.41	67.43±1.61	71.76±1.75	0.75	75.69±1.50	67.54±5.77	81.33±6.37
PSR	67.29±3.33	64.27±4.99	69.33±6.42	0.69	71.48±1.93	65.47±7.72	75.67±5.07
VTC+PSR	71.40±0.12	63.83±4.31	76.48±2.80	0.76	79.11±1.32	65.24±5.07	88.67±3.70
R	65.46±3.20	60.26±4.54	69.00±7.79	0.63	68.01±2.62	57.75±7.29	75.11±8.20
GDAM	66.14±1.12	59.65±3.90	70.55±4.01	0.69	70.56±1.86	59.66±5.06	78.11±6.07
TEOF	61.44±0.59	45.90±1.58	71.98±1.61	0.60	69.72±2.10	45.21±3.30	86.67±1.88
MFCC	78.79±1.53	79.61±3.36	78.25±2.98	0.87	83.51±1.97	80.58±4.31	85.55±3.83
Aco+Noi+MFCC	81.69±0.82	80.19±3.96	82.72±3.25	0.90	84.62±0.86	81.13±6.26	87.08±4.24
NLD+Entropy	64.44±2.14	62.18±2.35	65.98±4.09	0.67	69.25±2.19	61.53±3.98	74.58±2.98

Table 3.8: Hypernasality detection using VTC and PSR features for /i/ vowel.

	Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**AUROC**	**Acc (%)**	**Sen (%)**	**Spe (%)**
VTC	71.87±2.27	65.36±4.38	78.98±1.36	0.75	75.11±4.37	64.95±7.76	86.63±1.94
PSR	73.35±0.90	74.66±2.11	71.90±0.66	0.81	80.54±0.62	79.66±3.95	81.70±3.60
VTC+PSR	76.57±0.87	77.54±1.33	75.52±2.73	0.83	82.74±1.69	83.97±0.67	81.03±2.98
R	68.86±2.01	66.92±3.52	71.04±1.25	0.74	69.38±3.12	68.18±3.21	70.62±3.92
GDAM	62.78±1.99	46.08±7.51	80.98±6.12	0.67	63.70±2.32	42.37±8.81	87.74±7.39
TEOF	61.50±4.01	38.51±7.55	86.28±3.04	0.70	60.38±7.45	32.29±15.56	91.68±4.74
MFCC	84.68±0.56	83.74±1.98	85.29±1.27	0.94	89.87±0.86	87.18±2.64	92.76±1.85
Aco+Noi+MFCC	85.63±0.62	86.12±1.00	85.01±1.48	0.95	88.87±0.44	88.93±2.62	88.55±3.16
NLD+Entropy	77.12±0.56	77.09±1.94	77.01±1.60	0.86	81.04±0.48	80.83±2.92	81.03±2.45

Table 3.9: Hypernasality detection using VTC and PSR features for /u/ vowel.

	Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**AUROC**	**Acc (%)**	**Sen (%)**	**Spe (%)**
VTC	75.94±1.26	74.78±3.42	76.88±1.62	0.84	77.63±1.45	74.00±3.47	80.53±1.79
PSR	71.34±1.75	74.12±4.17	69.02±0.78	0.80	78.37±2.97	85.83±6.69	72.40±1.53
VTC+PSR	80.35±1.38	84.83±3.35	76.62±0.48	0.89	84.44±2.14	88.00±5.97	81.60±1.46
R	71.49±2.79	60.78±6.82	80.21±2.64	0.71	72.67±3.81	56.67±7.55	85.47±3.18
GDAM	66.04±3.42	66.11±5.63	65.91±1.99	0.67	67.11±4.23	64.67±7.13	69.07±2.39
TEOF	49.29±1.90	38.01±3.93	58.59±2.42	0.53	49.85±2.90	35.17±6.05	61.60±2.93
MFCC	85.09±2.16	91.14±5.67	80.08±1.90	0.95	87.19±1.19	93.00±2.86	82.53±1.79
Aco+Noi+MFCC	83.35±2.01	89.79±4.32	78.11±1.38	0.92	85.41±2.20	90.49±5.42	81.33±1.70
NLD+Entropy	60.16±1.76	56.64±2.71	63.10±4.19	0.64	66.90±2.90	59.03±4.47	73.20±4.56

of VTC and PSR features are better compared to the R, GDAM and TEOF features for each vowel. Another advantage of VTC and PSR features is that the features do not involve the low-pass filtering of the speech signal to capture the nasality evidence, so they do not require a predefined cutoff frequency, unlike these baseline features. The dependency of these baseline features on a suitable cutoff frequency may be the reason for their poor performance. This result shows that the individual VTC and PSR features can be a better choice for hypernasality detection compared to these baseline features. The individual performances of VTC and PSR features are also better compared to (NLD+Entropy) feature for /a/ and /u/ vowels, but it is poor for /i/ vowels. This shows that the (NLD+Entropy) feature is not properly capturing the nasality evidence present in hypernasal vowels. Although the better performance of individual VTC and PSR features compared to R, GDAM, TEOF features, their performance accuracy are low, and the best accuracy of only 80.54%, at phoneme level, is obtained using PSR feature for /i/ vowel. However, when VTC and PSR features are combined to form a two-dimensional (VTC+PSR) feature the performance increases. The improvement in performance shows that the VTC and PSR features are complementary to each other. McNemars statistical test also shows that the increment in the performance for combined (VTC+PSR) feature compared to individual VTC and PSR features is statistically significant ($p < 0.001$). The combined (VTC+PSR) feature gives an accuracy of 79.11%, 82.74%, and 84.44% at the phoneme level for /a/, /i/, and /u/ vowels, respectively. The accuracy of combined (VTC+PSR) feature is better compared to the individual performance of VTC, PSR and baseline R, GDAM, TEOF and (NLD+entropy) features. McNemars statistical test is also performed to compare the performance of the (VTC+PSR) feature with the best performing feature among these baseline features for each vowel, and it is found statistically significant ($p < 0.001$).

The performance comparison of (VTC+PSR) feature with MFCC and (Aco+Noi+MFCC) baseline features show that the performance of (VTC+PSR) feature is poor compared to these baseline features. The (Aco+Noi+MFCC) feature perform best for /a/ vowel, whereas the MFCC feature perform best for /i/ and /u/ vowels. For /a/ vowel also, the performance of MFCC feature is close to (Aco+Noi+MFCC) feature. The better performance of the MFCC feature for all three vowels may be because of its ability to model the whole spectral envelope of vowels which contains the vocal tract shape information. The shape of vocal tract differs for normal and hypernasal speech as the coupling of nasal tract with the oral tract happens during the production of hypernasal speech.

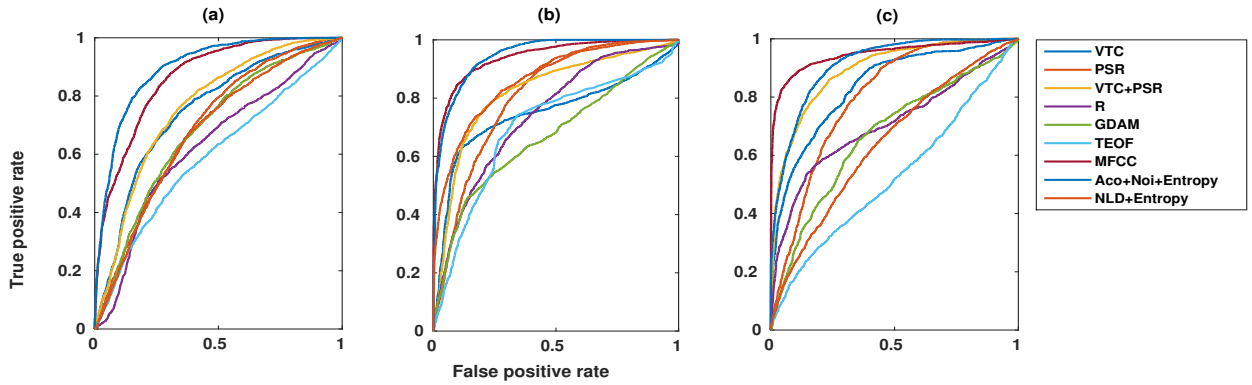

Figure 3.12: ROC curve for different features for different vowels. (a) for /a/vowel (b) for /i/ vowel and (c) for /u/vowel.

Fig. 3.12 (a)-(c) show the Receiver Operating Characteristic (ROC) curve corresponding to each feature for /a/, /i/ and /u/ vowels respectively. The curve is drawn for set 5 of the training-testing sets given in Table 3.6 and their Area Under Receiver Operating Characteristic (AUROC) curves is given in Table 3.7, Table 3.8, and Table 3.9 for /a/, /i/ and /u/ vowels respectively. The AUROC is higher for (VTC+PSR) feature compared to R, GDAM, TEOF and (NLD+entropy) baseline features for each vowel but it is lower compared to MFCC and (Aco+Noi+MFCC) Features.

3.6 Summary

This chapter of the thesis proposes two temporal features, VTC and PSR, for hypernasality detection. The VTC feature is extracted from the speech signal to capture the prominence of lower-frequencies in the signal. The value of VTC feature is found higher for hypernasal speech compared to normal speech. The PSR feature is extracted from the LP residual signal of the speech to capture the residual signal characteristics around epoch locations. The value of PSR feature is found lower for hypernasal speech compared to normal speech. The box plots for VTC and PSR features show their distinctive nature for normal and hypernasal speech. The result of ANOVA test shows the statistical significance of both the features for normal and hypernasal speech. The performance of two-dimensional (VTC+PSR) feature for normal and hypernasal speech classification using the SVM classifier is found better compared to the baseline R, GDAM TEOF and (NLD+entropy) features for each vowel. The comparison of performance for (VTC+PSR) feature with the MFCC and (Aco+Noi+MFCC) features shows that the MFCC and (Aco+Noi+MFCC) feature performance bet-

ter compared to (VTC+PSR) feature. The better performance of MFCC feature may be because of modeling of the spectral envelope of a vowel by cepstral coefficients.

In this chapter, the effect of the presence of nasal formant and antiformant pairs in hypernasal vowels spectrum is captured indirectly using the temporal features VTC and PSR. As nasal formant and antiformant pairs affect the strength of harmonics directly, so in the next chapter, their effects are directly captured by computing the strength of harmonics using the sinusoidal model of speech for hypernasality detection.

4

Hypernasality detection using sinusoidal model-based features

Publications

- **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Sinusoidal model-based hypernasality detection in cleft palate speech using CVCV sequence," *Speech Communication*(2020).

Contents

Overview

The presence of formant and antiformant pairs in the hypernasal vowel spectrum affects the strength of harmonics. The nasal formants increases whereas the antiformants decrease the strength of harmonics around its location of addition. This may give the difference in the pattern of harmonics strength of normal and hypernasal vowels. So in this chapter, three features namely, normalized harmonics amplitude (NHA), harmonics amplitude ratio (HAR) and prominent harmonics frequency (PHF) which are based on the strength of harmonics are used to capture the spectral difference between normal and hypernasal vowels spectrum for their classification. The NHA feature captures the relative strength of harmonics with respect to their maximum strength. The HAR feature is the ratio of strength of a harmonic to the strength of its preceding harmonic. The PHF feature is based on the region of prominent harmonics in the spectrum. The maximum strength of harmonic, pattern of harmonics strength with respect to neighbouring harmonics strength and region of prominent harmonics may differ in hypernasal vowel compared to normal vowel due to the presence of nasal formant and antiformant pairs. So these features may be discriminating and could be informative for classification of normal and hypernasal speech.

4.1 Introduction

In Chapter 3, the effect of presence of nasal formant and antiformant pairs in the spectrum of hypernasal vowels is captured using temporal features VTC and PSR extracted from the speech signal and its residual signal respectively. However, the effect can also be captured in the spectral domain from the vowel spectrum itself. In literature, it has been done either by formant analysis or gross spectrum shape analysis. In formant analysis, a high-resolution spectrum such as group delay spectrum [26] has been used to resolve the nasal and the oral formants and based on their strength feature has been proposed for hypernasality detection. But due to the high fundamental frequency of children's speech and the presence of nasal antiformants, the formants in hypernasal speech may not be prominent enough to be easily resolved. Hence, the peaks-picked by peak picking algorithm may not actually represent the original formants. It is also reported that the formant analysis cannot be used for clinical application because it is not real-time [40]. On the other hand, the gross spectrum shape analysis involves modeling of the envelope of the vowels spectrum with the cepstral coefficients. Hence, features such as MFCC and LPCC have been used in the literature for

the detection of hypernasality [32,33]. But, the studies in [45,89] show that these features are unable to adequately smooth out the effect of pitch harmonic in the low-frequency region of the spectrum for the children's speech having a high fundamental frequency. So, the features give a high variance for the higher coefficients. Due to the inadequate smoothing, these features may not be able to capture the nasality evidence present in the low-frequency region. Because of the above-mentioned issues with the formant analysis and gross spectrum shape analysis, another way of spectrum analysis to analyze the hypernasal vowels is explored in this chapter. In this method, the strength of harmonics in the spectrum is analyzed. The pitch and its harmonics are the key attributes of a vowel spectrum and their strength gets directly affected due to the addition of nasal formant and antiformant pairs in the spectrum. The strength of harmonics gets enhanced around the location of addition of nasal formant and it gets diminished around the location of addition of nasal antiformant. This effect of nasal formant and antiformant pairs on harmonics strength is also applicable for children's speech with the high fundamental frequency. This harmonics strength analysis of hypernasal vowels is also motivated with the study in [53] which suggests that an objective measure for the detection of hypernasality can be developed for clinical purposes based on the harmonics-intensity-measurement-procedure because the basic difference between normal and hypernasal vowels is found in the strength of harmonic intensities. These differences in the form of harmonic strength are becomes significant for the high pitch children speech.

This work explores the sinusoidal model of speech to compute the strength of harmonics from the vowels spectrum. Three features namely, normalized harmonics amplitude (NHA), harmonics amplitude ratio (HAR) and prominent harmonics frequency (PHF) are used for hypernasality detection. Features are based on the computed harmonics strength and capture the spectral difference between normal and hypernasal vowels for their classification. The NHA feature captures the relative strength of harmonics with respect to their maximum strength. The HAR feature is the ratio of strength of a harmonic to the strength of its preceding harmonic. So the feature captures the relative harmonics strength. The PHF feature is based on the region of prominent harmonics in the spectrum and captures the frequency of prominent harmonics. The hypothesis is that as the maximum strength of harmonic, pattern of harmonics strength with respect to neighbouring harmonics strength and region of prominent harmonics may differ in hypernasal vowels compared to normal vowels due to the presence of nasal formant and antiformant pairs. Hence, NHA, HAR, and PHF features may be

discriminating for normal and hypernasal vowels. Also, the proposed features are simply computed from the DFT spectrum rather than a high-resolution spectrum. So these features could be used for the classification of normal and hypernasal vowels using the SVM classifier.

The rest of this chapter is organized as follows: Section 4.2 describes the sinusoidal model of speech and sinusoidal parameters for hypernasality detection. Section 4.3 describes the sinusoidal model-based features. Section 4.4 describes the discriminative capability of feature dimensions. Section 4.5 describes the experimental results and Section 4.6 summarizes the work and provides the conclusion of the chapter.

4.2 Sinusoidal model of speech

The spectral characteristics of hypernasal vowels are different from normal vowels due to the addition of nasal formant and antiformant pairs as explained in Section 3.3.1. The presence of nasal formant and antiformant pairs affects the strength of harmonics in the hypernasal vowels spectrum. Since the sinusoidal model for speech waveform leads to an analysis technique in term of amplitudes, frequencies, and phases of the harmonics component, this section of the chapter discusses the sinusoidal model-based features for hypernasality detection.

The speech signal can be assumed as the output of the convolution of the glottal excitation signal with the time-varying linear filter which models the resonant characteristics of the vocal tract. According to the sinusoidal model of speech, a glottal excitation signal can be represented in terms of a sum of sine waves of arbitrary amplitudes, frequencies, and phases [141] and the speech signal divided into M frames can be given as [142],

$$s[n] = \sum_{k=1}^{M} \hat{s}[n - kN], \tag{4.1}$$

where k is the frame index, and N is the length of the frame. $\hat{s}[n]$ is a sum of sinusoids given by,

$$\hat{s}[n] = \sum_{j=1}^{L} A_j^k cos(2\pi f_j^k \frac{n}{F_s} + \phi_j^k), \tag{4.2}$$

where F_s is the sampling frequency of $s[n]$ and L is the order of the sinusoidal model. A_j^k, and ϕ_j^k are the amplitude and phase of j^{th} sine wave along the frequency track f_j^k which lies in the interval $0 \leq f_j^k \geq F_s/2$. The A^k, f^k and ϕ^k are considered as the three parameters of a sinusoid in the sinusoidal model for time frame k.

Figure 4.1: Block diagram representing steps of extraction of harmonic amplitude and frequency.

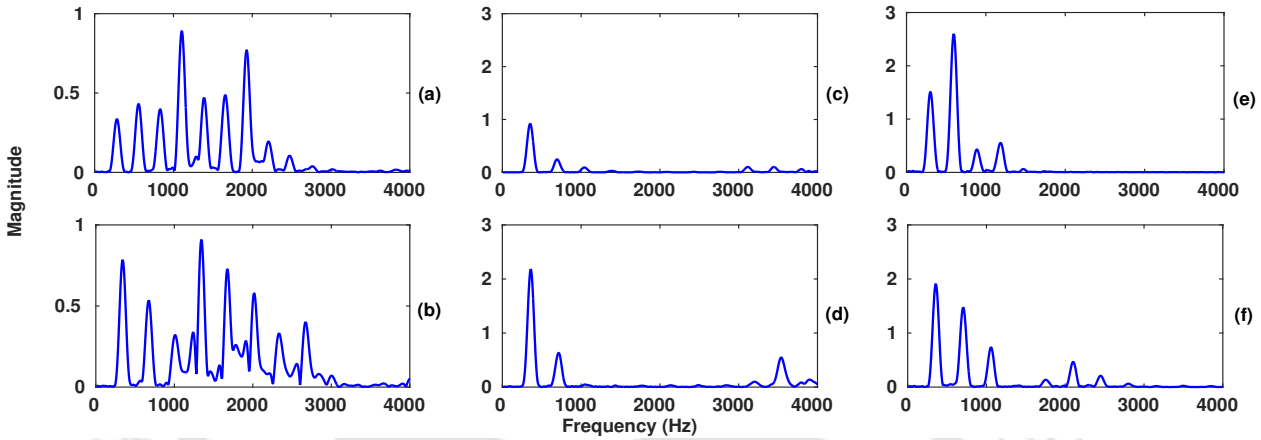

Figure 4.2: Illustration of difference in DFT spectrum of normal and hypernasal speech. (a)-(b) are the spectrum of normal and hypernasal /a/ vowels, respectively, (c)-(d) are the spectrum of normal and hypernasal /i/ vowels, respectively and (e)-(f) are the spectrum of normal and hypernasal /u/ vowels, respectively.

4.2.1 Sinusoidal parameters for hypernasality detection

Among the three parameters (A^k, f^k and ϕ^k) of a sinusoid in the sinusoidal model, the amplitude and frequency of sinusoids are directly affected by the addition of nasal formant and the antiformant pairs in hypernasal speech spectrum. The strength of harmonics increases around the frequency location of the addition of nasal formants, and it decreases around the frequency location of the addition of antiformants. The addition of nasal formants and antiformant pairs also changes the frequency of prominent harmonics in the spectrum. Hence, the amplitude and frequency of sinusoids can be used for hypernasality detection.

The block diagram shown in Figure 4.1 gives the steps followed to extract amplitude and frequency parameters of sinusoids. The pre-processing stage divides the speech signals into a sequence of overlapping frames, and each frame is multiplied with the Hamming window to remove the signal discontinuities at the ends of each frame. In the next step, the discrete Fourier transform (DFT) of each frame is performed. Figure 4.2 shows the DFT spectrum of normal and hypernasal speech. Figure 4.2 (a)-(b) are the spectrum of normal and hypernasal /a/ vowel respectively. It can be observed that the hypernasal /a/ vowel has higher magnitude harmonics in low-frequency below 500 Hz

and lower magnitude harmonics around F_1 at 800 Hz compared to the normal /a/ vowel. Figure 4.2 (c)-(d) are the spectrum of normal and hypernasal /i/ vowel respectively. In this case, the hypernasal /i/ vowel has higher magnitude harmonics in low-frequency at F_1 around 300 Hz and also in higher frequency above 3000 Hz compared to normal /i/ vowel. Figure 4.2 (e)-(f) are the spectrum of normal and hypernasal /u/ vowel respectively. In this case also the higher magnitude harmonics are present in the low-frequency at F_1 around 400 Hz but the magnitude of harmonics around 800 Hz gets reduced compared to normal /u/ vowel. Hence, it can be observed from the Figure 4.2 that the magnitude of harmonics and the frequency of prominent harmonics both get modified in hypernasal speech compared to normal speech due to nasal formants and antiformants pairs.

In the DFT spectrum, the sinusoids are present in the form of the fundamental frequency and its harmonics, whose amplitude and frequency can be obtained using the peak-picking algorithm. The steps for obtaining the amplitude and frequency of L harmonics in the DFT spectrum are as follows:

Step1: Apply the peak-picking algorithm to obtain all the peaks present in the spectrum.

Step2: Compute the fundamental frequency (f_0) of the frame using the instantaneous fundamental frequencies obtained from zero-frequency filtered signal proposed in [143]. The instantaneous fundamental frequencies within the frame are averaged to obtain the f_0 for the frame.

Step3: Search for the highest peak among all peaks obtained in Step1 in the interval of $[f_0 - f_0/3, f_0 + f_0/3]$ [144]. Consider the frequency of this highest peak as the first harmonic peak frequency denoted by f_1. The interval is taken in such a way so that no two neighboring intervals overlap with each other.

Step4: Use f_1 to obtain the remaining $L-1$ harmonic peaks. This is done by searching the highest peak among all peaks obtained in step 1 in the interval of $[i \times f_0 - f_0/3, i \times f_0 + f_0/3]$, where $i = 2, 3...L$. Denote these harmonic peak frequencies by $[f_2, f_3, ..., f_L]$.

Step5: Compute the amplitudes of all L harmonics at the $[f_1, f_2, ..., f_L]$ frequencies. Denote the amplitudes by $[A_1, A_2, ..., A_L]$. These amplitude values and corresponding frequencies are parameters of sinusoids in the sinusoid model.

Figure 4.2 shows the effect of the presence of nasal formant and antiformant in the spectrum of hypernasal vowels on the strength of harmonics for a single frame. The effect can be better represented

by the error bars for the strength of first 10 harmonics of all frames corresponding to normal and hypernasal speech present in the database discussed in Section 3.2. The value of L is taken 10 in this work that will cover the lower frequency range (below 3000 Hz, considering f_0 in the range of 250 to 275 Hz) of the spectrum where nasality evidence is mostly present. Fig. 4.3 shows these error bar plots with mean and std. The error bar plots are shown for the normal vowel at the $1, 3, ..., 19$ indices and for the hypernasal vowel at the $2, 4, ..., 20$ indices. Fig. 4.3 (a) is shown for /a/ vowel, (b) for /i/ vowel, and (c) for /u/ vowel. It can be observed from Fig. 4.3 that the mean and std of harmonics is greater for hypernasal vowel compared to normal vowel. So, the Fig. 4.3 validates the argument presented in Fig. 4.2. Hence, features based on harmonic strength and their frequencies can be used for normal and hypernasal vowel classification.

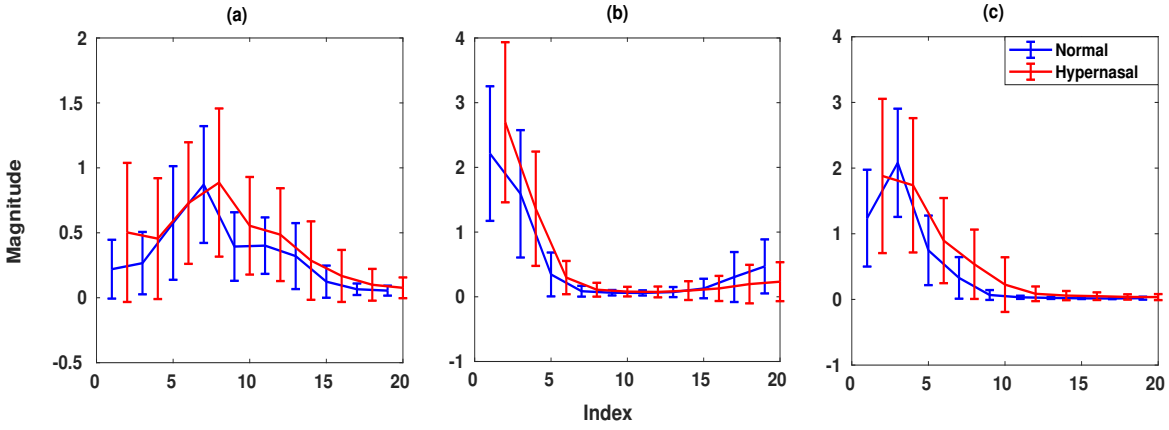

Figure 4.3: Error bar plots (a) for /a/ vowel (b) for /i/ vowel (c) for /u/ vowel show the difference in the strength of first 10 harmonics of normal and hypernasal vowels present in the entire database. Indices $1, 3, ..., 19$ are for normal vowel harmonics, whereas the indices $2, 4, ..., 20$ are for hypernasal vowel harmonics.

4.3 Sinusoidal model-based features

The amplitudes and corresponding frequencies of harmonics in the spectrum of normal and hypernasal speech are different due to nasal formants and anti-formants. Hence, features are extracted from the amplitude and frequency of harmonics for the detection of hypernasal speech. The amplitude and frequency parameters of $L + 1$ harmonics are extracted from the normalized spectrum using the procedure discussed in the previous subsection. The spectrum of each frame is divided by the total energy of that frame to compute the normalized spectrum. The various features are as follows.

4.3.1 Normalized harmonics amplitude feature

The harmonics amplitude $[A_1, A_2, ..., A_L, A_{L+1}]$ are divided by the maximum amplitude value among them to normalize the amplitude values. This type of normalization will give one value equal to 1 corresponding to the maximum amplitude value. All the normalized harmonic amplitudes (NHA) except the value equal to 1 are taken as the NHA feature and its total dimensions are 10. The NHA feature gives the relative amplitude of harmonics with respect to the maximum amplitude. As the maximum amplitude for hypernasal vowels is different compared to normal vowels, the nature of NHA feature may also differ for hypernasal and normal vowels. Fig. 4.4 shows the nature of three different dimensions of NHA feature for normal and hypernasal /a/, /i/ and /u/ vowels, respectively. Fig. 4.4 (a)-(c) are shown for /a/ vowel, Fig. 4.4 (d)-(f) are shown for /i/ vowel, and Fig. 4.4 (g)-(i) are shown for /u/ vowel. The NHA feature is plotted for 200 frames of normal vowel and 200 frames of hypernasal vowel, respectively. It can be observed from Fig. 4.4 (a), (d), (g), the nature of NHA feature is quite different for normal and hypernasal vowels for a particular dimension, and this difference reduces for another dimension as shown in Fig. 4.4 (b), (e), (h). There is also some other dimension for which the nature of NHA feature is quite similar as shown in Fig. 4.4 (c), (f), (i).

The harmonics amplitude $[A_1, A_2, ..., A_L, A_{L+1}]$ are divided by the maximum amplitude value among them to normalize the amplitude values. This type of normalization will give one value equal to 1 corresponding to the maximum amplitude value. All the normalized harmonics amplitude (NHA) except the value equal to 1 are taken as the NHA feature. This feature gives the relative amplitude w.r.t. the maximum amplitude.

4.3.2 Harmonics amplitude ratio feature

HAR feature is the successive ratio of harmonics amplitude $[A_1, A_2, ..., A_L, A_{L+1}]$. It is the ratio of a harmonic magnitude with its previous harmonic magnitude. It can be represented by $[\frac{A_2}{A_1}, \frac{A_3}{A_2}, ..., \frac{A_L}{A_{L-1}}, \frac{A_{L+1}}{A_L}]$. The total dimensions of HAR feature are $L = 10$. As the pattern of harmonic amplitudes is different for normal and hypernasal vowels, the nature of HAR feature may also differ for hypernasal and normal vowels. Fig. 4.5 shows the nature of three different dimensions of HAR feature for normal and hypernasal /a/, /i/ and /u/ vowels, respectively. Fig. 4.5 (a)-(c) is shown for /a/ vowel, Fig. 4.5 (d)-(f) is shown for /i/ vowel, and Fig. 4.5 (g)-(i) is shown for /u/ vowel. The HAR feature is also plotted for 200 frames of normal vowel and 200 frames of hypernasal

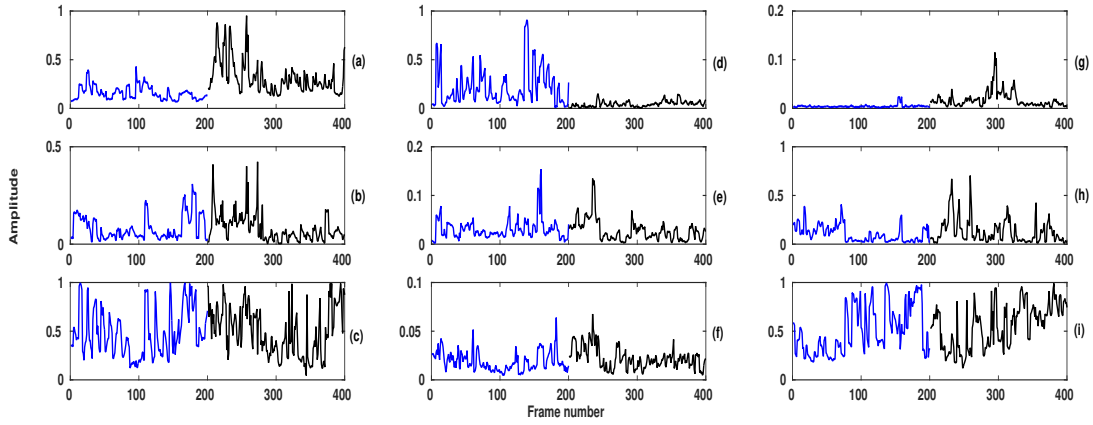

Figure 4.4: Illustration of nature of NHA feature in normal and hypernasal speech. (a)-(c) is for vowel /a/, (d)-(f) for vowel /i/ and (g)-(i) is for vowel /u/.

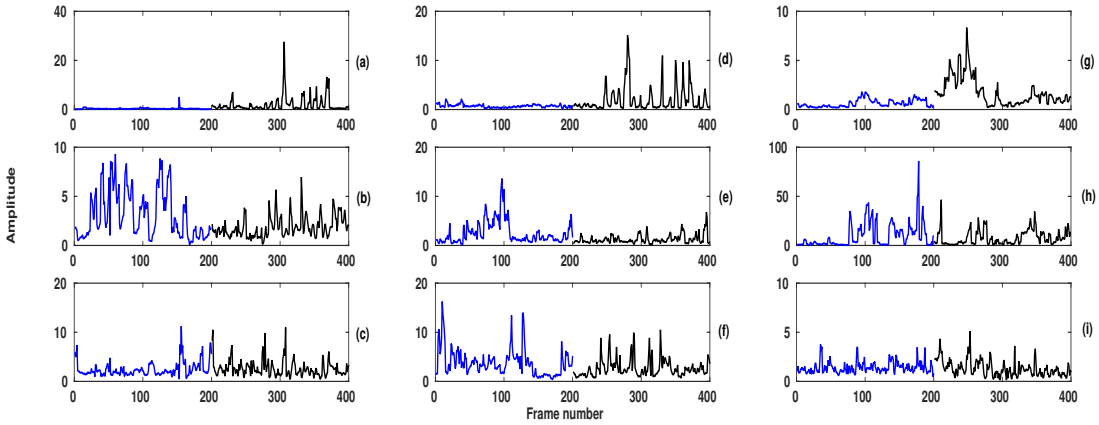

Figure 4.5: Illustration of nature of HAR feature in normal and hypernasal speech. (a)-(c) is for vowel /a/, (d)-(f) for vowel /i/ and (g)-(i) is for vowel /u/.

vowel, respectively. Like the NHA feature, it can be observed from Fig. 4.5 (a), (d), (g), the nature of HAR feature is quite different for normal and hypernasal vowels for a particular dimension, and this difference reduces for another dimension as shown in Fig. 4.5 (b), (e), (h). There is also some other dimension for which the nature of HAR feature is quite similar as shown in Fig. 4.5 (c), (f), (i).

4.3.3 Prominent harmonics frequency feature

The PHF feature is the frequency of L most prominent harmonics. It is computed by sorting the harmonic amplitudes $[A_1, A_2, ..., A_L, A_{L+1}]$ in descending order and choosing the L most prominent harmonic amplitudes from them. The L frequencies corresponding to the chosen harmonics can be obtained from $[f_1, f_2, ..., f_L, f_{L+1}]$ frequencies and they are called PHF feature. The value of L is

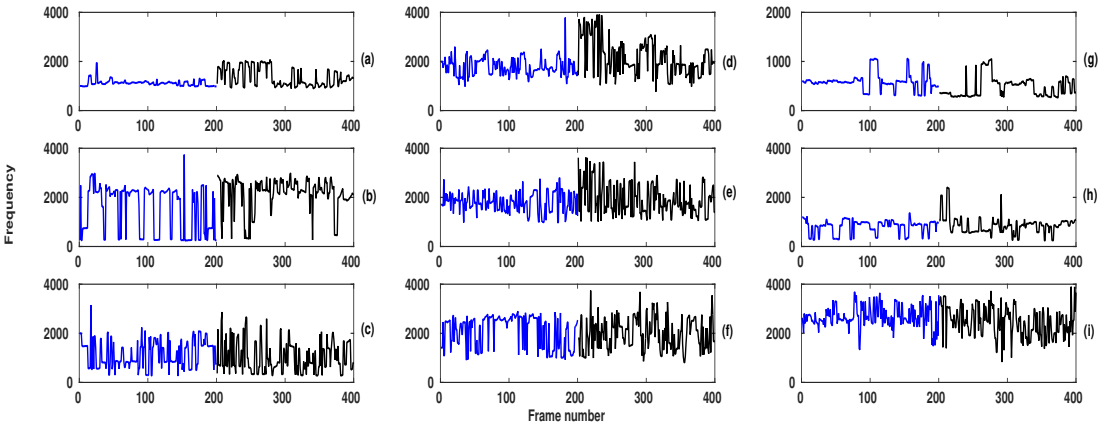

Figure 4.6: Illustration of nature of PHF feature in normal and hypernasal speech. (a)-(c) is for vowel /a/, (d)-(f) for vowel /i/ and (g)-(i) is for vowel /u/.

taken as 10, so the total dimensions of the HAR feature are 10. Fig. 4.6 shows the nature of three different dimensions of PHF feature for normal and hypernasal /a/, /i/ and /u/ vowels, respectively. Fig. 4.6 (a)-(c) are shown for /a/ vowel, Fig. 4.5 (d)-(f) are shown for /i/ vowel, and Fig. 4.6 (g)-(i) are shown for /u/ vowel. The PHF feature is also plotted for 200 frames of normal vowel and 200 frames of hypernasal vowel, respectively. Like the NHA and HAR features, it can be observed from Fig. 4.6 (a), (d), (g), the nature of PHF feature is quite different for normal and hypernasal vowels for a particular dimension, and this difference reduces for another dimension as shown in Fig. 4.6 (b), (e), (h). There is also some other dimension for which the nature of PHF feature is quite similar as shown in Fig. 4.6 (c), (f), (i). From Fig. 4.4, Fig. 4.5 and Fig. 4.6, it is clear that out of 10 dimensions of NHA, HAR, and PHF features, there are some dimensions of features which are having quite discriminative values for normal and hypernasal vowels, whereas there are also some dimensions for which the values are quite similar. The values of features for rest dimensions lie between these extrema. Hence, it would be better to find the discriminative capability of each dimension of each feature for normal and hypernasal vowels.

4.4 Discriminative capability of feature dimensions

The discriminative capability of a particular dimension of a feature for normal and hypernasal vowels can be estimated by measuring the statistical dependency (SD) between features and class labels. The SD measure is used to find whether the values of a feature are dependent on the associated

class labels, or whether the two simply co-occur by chance. The measure has been used for feature selection for the classification of speaker likability, intelligibility, and personality traits [145]. The measure involves quantization of the feature values into one of the feature-specific quantization levels. These levels are determined adaptively in such a way that each bin will contain roughly an equal amount of samples over the entire data set. The statistical dependence between the discretized feature values y and the class labels z is computed according to the formula,

$$SD = \sum_{y \epsilon Y} \sum_{z \epsilon Z} P(y,z) \frac{P(y,z)}{P(y)P(z)}, \tag{4.3}$$

A large SD value indicates a high dependency between the feature values and the class labels. The measure is similar to the mutual information (MI) given by,

$$MI = \sum_{y \epsilon Y} \sum_{z \epsilon Z} P(y,z) \log \left[\frac{P(y,z)}{P(y)P(z))} \right], \tag{4.4}$$

MI is also a statistical measure and has been used for feature selection [146]. Table 4.1, Table 4.2 and Table 4.3 show the discriminative capability of each dimension of NHA, HAR, and PHF features for /a/, /i/ and /u/ vowels, respectively. The SD measures arranged in descending order and corresponding feature dimensions are shown in these tables. The number of quantization levels considered is 12 as suggested in [145]. Higher the value of SD measure, better the discriminative capability of that feature dimension. Fig. 4.7, Fig. 4.8, and Fig. 4.9 show the normalized histogram of three dimensions of NHA, HAR, and PHF features, respectively, with higher SD measure for normal and hypernasal vowels. Part (a)-(c) of Fig. 4.7, Fig. 4.8 and Fig. 4.9 are plotted for vowel /a/, (d)-(f) for vowel /i/, and (g)-(i) for vowels /u/. The histograms show the discriminative nature of features for normal and hypernasal vowels and supports the point that feature dimension with higher SD measure has a better discriminative capability. Table 4.4, Table 4.5, and Table 4.6 show the mean and std of three dimensions of NHA, HAR, and PHF features, respectively, with higher SD measure for normal and hypernasal vowels. The mean of NHA, HAR, and PHF features is also different for normal and hypernasal vowels. This analysis demonstrates that the NHA, HAR, and PHF features are discriminative for normal and hypernasal vowels, and the feature dimensions with higher SD measures are more discriminative. So these dimensions of features can be used for normal and hypernasal vowels classification.

Table 4.1: NHA feature dimensions and corresponding SD measure for vowels /a/, /i/ and /u/.

/a/		/i/		/u/	
Dimension	SD measure	Dimension	SD measure	Dimension	SD measure
1	0.2332	10	0.2557	7	0.1687
2	0.1088	9	0.1306	6	0.1351
3	0.0575	8	0.0932	8	0.1290
4	0.0382	1	0.0931	5	0.1270
8	0.0330	7	0.0698	4	0.1201
9	0.0292	2	0.0513	3	0.0574
10	0.0228	4	0.0253	9	0.0525
6	0.0221	6	0.0166	2	0.0180
7	0.0098	5	0.0163	10	0.0135
5	0.0082	3	0.0086	1	0.0078

Table 4.2: HAR feature dimensions and corresponding SD measure for vowels /a/, /i/ and /u/.

/a/		/i/		/u/	
Dimension	SD measure	Dimension	SD measure	Dimension	SD measure
3	0.0740	7	0.1129	1	0.1765
1	0.0582	8	0.1116	10	0.0532
4	0.0576	1	0.0848	2	0.0490
7	0.0474	2	0.0678	9	0.0441
6	0.0467	9	0.0548	4	0.0379
2	0.0451	3	0.0504	5	0.0345
5	0.0302	6	0.0405	3	0.0321
9	0.0167	10	0.0310	8	0.0282
10	0.0147	5	0.0230	6	0.0266
8	0.0125	4	0.0080	7	0.0254

Table 4.3: PHF feature dimensions and corresponding SD measure for vowels /a/, /i/ and /u/.

/a/		/i/		/u/	
Dimension	SD measure	Dimension	SD measure	Dimension	SD measure
1	0.0658	4	0.1393	1	0.2079
6	0.0597	10	0.1213	2	0.1161
10	0.0520	2	0.1124	5	0.1033
8	0.0508	5	0.1056	6	0.0972
7	0.0501	6	0.1040	3	0.0912
9	0.0460	3	0.1029	4	0.0769
3	0.0450	7	0.0966	10	0.0587
2	0.0427	8	0.0926	7	0.0509
4	0.0406	9	0.0875	8	0.0369
5	0.0125	1	0.0615	9	0.0106

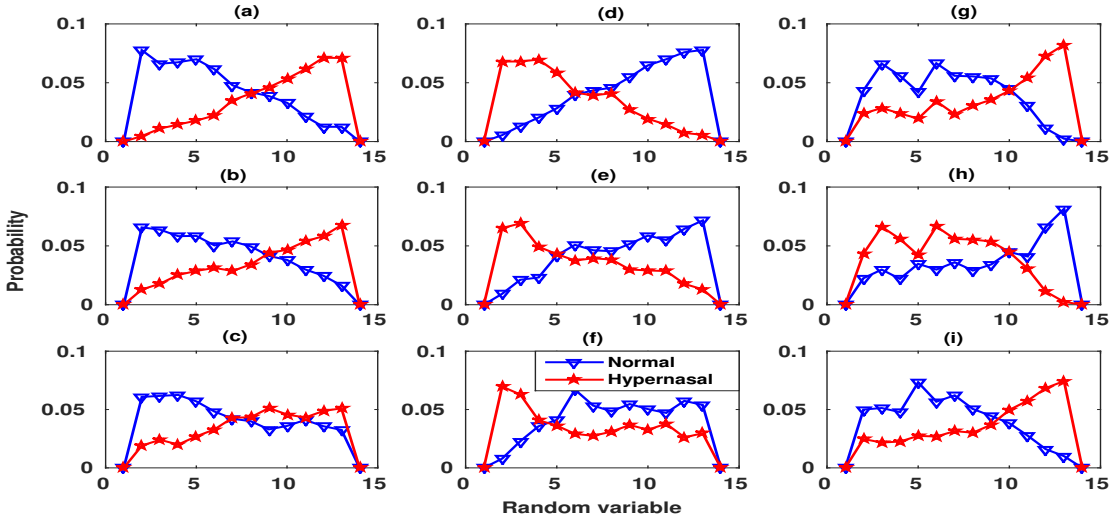

Figure 4.7: Normalized histogram of first three high SD measure dimension of NHA feature. (a)-(c) for vowel /a/, (d)-(f) for /i/ vowel and (g)-(i) for /u/ vowel.

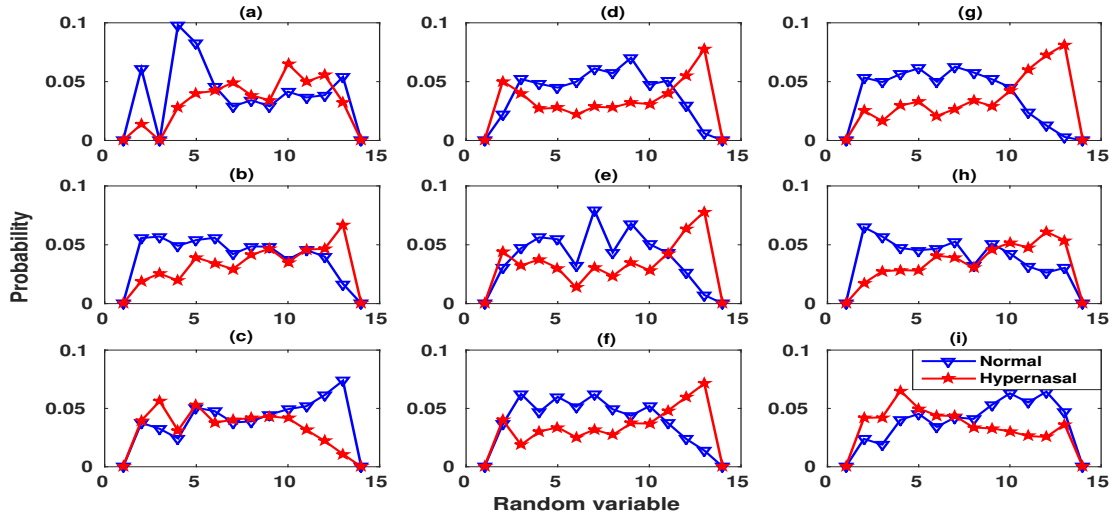

Figure 4.8: Normalized histogram of first three high SD measure dimension of HAR feature. (a)-(c) for vowel /a/, (d)-(f) for /i/ vowel and (g)-(i) for /u/ vowel.

Table 4.4: Mean (μ) and standard deviation (σ) for first three dimensions of NHA feature with higher SD measure for normal and hypernasal /a/, /i/ and /u/ vowels.

/a/		/i/		/u/	
Normal	Hypernasal	Normal	Hypernasal	Normal	Hypernasal
$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$
0.20±0.13	0.38±0.21	0.26±0.21	0.08±0.09	0.01±0.01	0.02±0.02
0.24±0.14	0.36±0.20	0.21±0.21	0.09±0.13	0.01±0.01	0.02±0.03
0.44±0.23	0.55±0.22	0.12±0.17	0.08±0.15	0.01±0.01	0.02±0.02

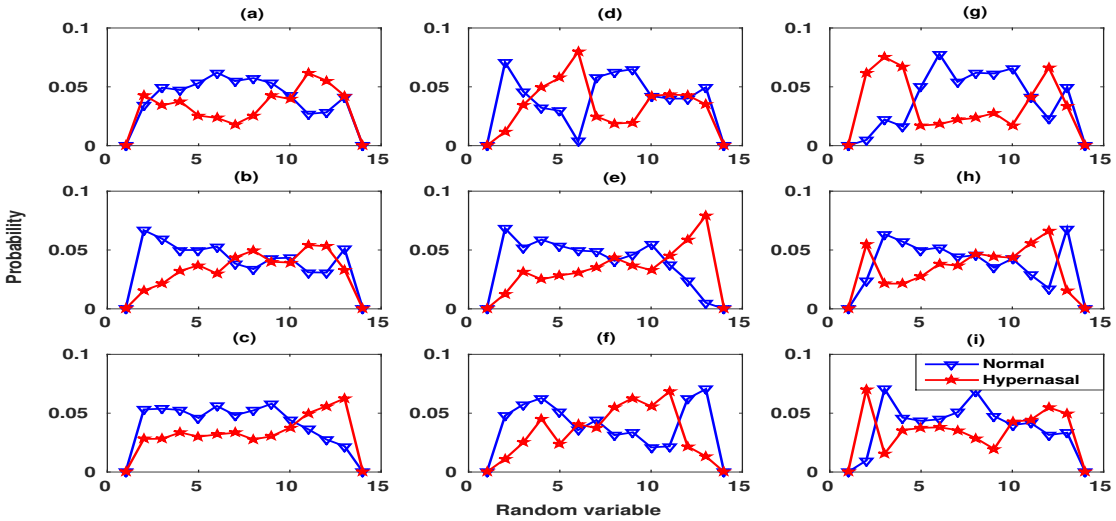

Figure 4.9: Normalized histogram of first three high SD measure dimension of PHF feature. (a)-(c) for vowel /a/, (d)-(f) for /i/ vowel and (g)-(i) for /u/ vowel.

Table 4.5: Mean (μ) and standard deviation (σ) for first three dimensions of HAR feature with higher SD measure for normal and hypernasal /a/, /i/ and /u/ vowels.

/a/		/i/		/u/	
Normal	Hypernasal	Normal	Hypernasal	Normal	Hypernasal
$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$
1.07±1.97	1.11±1.29	0.73±0.43	1.37±1.70	0.65±0.49	1.50±1.33
0.89±0.48	1.24±0.85	0.61±0.41	1.23±1.63	1.27±1.73	1.55±1.20
3.49±3.71	2.07±1.63	1.92±1.42	3.25±2.91	5.82±8.11	4.59±8.30

Table 4.6: Mean (μ) and standard deviation (σ) for first three dimensions of PHF feature with higher SD measure for normal and hypernasal /a/, /i/ and /u/ vowels.

/a/		/i/		/u/	
Normal	Hypernasal	Normal	Hypernasal	Normal	Hypernasal
$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$
1107.90±251.90	1157.30±363.41	2260.80±1082.10	2134.70±1030.20	585.77±174.27	523.55±248.42
1215.20±914.64	1337.60±809.41	1677.30±413.21	2069.00 ±607.25	484.44±269.94	621.17±307.09
1552.40±598.53	2788.00±546.08	1140.60±1205.00	784.04±743.20	1725.20±727.92	1745.0±659.16

4.5 Experiments and results

In this section, experimental setup and result of normal vs. hypernasal vowel classification using individual NHA, HAR, PHF and combined (NHA+HAR+PHF) features are discussed. The result of the proposed features is compared with the results obtained using the VTC+PSR feature and baseline MFCC and Acc+Noi+MFCC features. The result is not compared with the R, GDAM, TEOF, and NLD+Entropy features because the VTC+PSR feature performs better compared to these feature as discussed in Chapter 3. The result is evaluated using the SVM classifier.

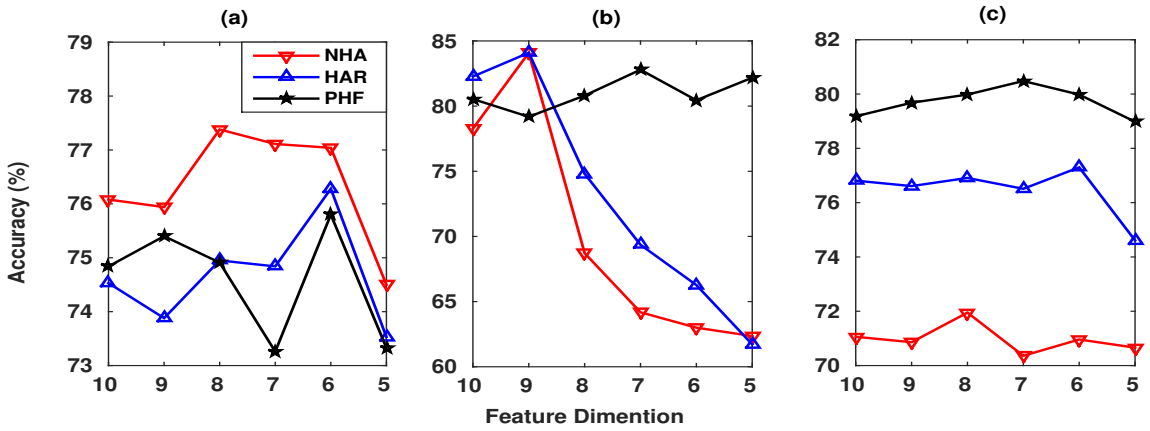

Figure 4.10: Accuracy corresponding to different features based on the number of feature dimensions for vowels. (a) /a/, (b) /i/ and (c) /u/.

4.5.1 Experiments

The proposed features NHA, HAR, and PHF are extracted from each frame of the speech signal. The frame size of 20 ms and frameshift of 10 ms is used for framing of speech. For each vowel, five classification experiments using five different training-testing sets are performed. The database, five different training-testing sets, and the RBF kernel parameters (c, γ) are the same that are used in Chapter 3. The experiments are performed using individual NHA, HAR, PHF features and a combination of these features (NHA+HAR+PHF) with their optimum dimensions. The SVM classification experiments are carried out to find the optimum dimension of features for each training-testing set. For that, SD measures are computed with 12 quantization levels and the feature dimensions are arranged in decreasing order of SD measure as shown in Table 4.1, Table 4.2, and Table 4.3. Repeated experiments are then done using all 10 dimensions, 9 higher SD value dimensions like that up to 5 higher SD value dimensions to find the classification result for each feature. The number of dimensions

that give the best result is reported as the classification result for that particular set. The optimum dimensions of (NHA+HAR+PHF) features are finalized by summing the optimum dimensions of individual NHA, HAR and PHF features. This can be understood from Fig. 4.10 which shows the number of dimensions from 10 to 5 on the horizontal axis and the accuracy for NHA, HAR, and PHF features on the vertical axis for the first training-testing set. Fig. 4.10 (a) is shown for vowel /a/, (b) is for vowel /i/ and (c) is for vowel /u/. For vowel /a/, the NHA feature gives the best accuracy by taking 8 higher SD dimensions of the feature, HAR feature and PHF feature give the best accuracy with 6 dimensions of the feature. Hence, the optimum dimension of the NHA feature is 8, for HAR and PHF feature it will be 6 and for combined (NHA+HAR+PHF) feature it will be $8 + 6 + 6 = 20$. Similarly for the vowels /i/ and /u/ the optimum dimensions of individual features and the combined feature can be obtained from Fig. 4.10. Similar repeated experiments are done to find the optimum dimension of features for other training-testing sets also.

4.5.2 Results

As in Chapter 3, the classification results of normal and hypernasal vowels are presented at the frame as well as the phoneme level. The results are presented in terms of accuracy (Acc), sensitivity (Sen) and specificity (Spe) parameters. Table 4.7, Table 4.8, and Table 4.9 show the classification performances for /a/, /i/ and /u/ vowels, respectively. Each table shows the performance for the NHA feature, HAR feature, PHF feature, combined (NHA+HAR+PHF) feature, VTC+PSR feature and baseline MFCC and Acc+Noi+MFCC features. The results are presented in the form of mean and std of results obtained for five training- testing sets for each feature.

Table 4.7: Hypernasality detection using NHA, HAR and PHF features for /a/ vowel.

	Frame level performance				**Vowel level performance**		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**AUROC**	**Acc (%)**	**Sen (%)**	**Spe (%)**
NHA	78.83±2.20	78.28±5.41	79.17±2.08	0.86	80.17±2.43	78.74±7.21	81.08±1.68
HAR	73.83±2.70	73.29±4.67	74.13±4.35	0.75	75.95±3.35	73.35±3.37	77.57±6.07
PHF	75.45±2.44	70.82±4.92	78.43±4.48	0.79	77.10±3.63	70.99±3.99	81.05±5.64
NHA+HAR+PHF	81.54±1.97	83.29±4.27	80.39±1.04	0.91	82.46±2.30	82.38±5.52	82.45±1.19
VTC+PSR	71.40±0.12	63.83±4.31	76.48±2.80	0.76	79.11±1.32	65.24±5.07	88.67±3.70
MFCC	78.79±1.53	79.61±3.36	78.25±2.98	0.87	83.51±1.97	80.58±4.31	85.55±3.83
Aco+Noi+MFCC	81.69±0.82	80.19±3.96	82.72±3.25	0.90	84.62±0.86	81.13±6.26	87.08±4.24

The results show that among NHA, HAR and PHF features, individually, the NHA feature performs better for /a/ vowel, HAR feature performs better for /i/ vowel and PHF feature perform better

Table 4.8: Hypernasality detection using NHA, HAR and PHF features for /*i*/ vowel.

	Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**AUROC**	**Acc (%)**	**Sen (%)**	**Spe (%)**
NHA	84.13±1.67	81.08±4.94	87.43±2.23	0.90	83.86±2.28	79.67±5.72	88.82±2.41
HAR	84.97±0.83	79.88±1.26	89.97±0.73	0.92	85.86±1.78	79.67±2.85	92.96±0.71
PHF	79.85±1.53	77.84±4.43	81.80±2.95	0.89	83.87±2.21	79.82±6.83	88.24±2.78
NHA+HAR+PHF	86.38±0.84	84.28±2.32	88.52±1.04	0.93	87.89±0.49	84.86±1.13	91.34±1.12
VTC+PSR	76.57±0.87	77.54±1.33	75.52±2.73	0.83	82.74±1.69	83.97±0.67	81.03±2.98
MFCC	84.68±0.56	83.74±1.98	85.29±1.27	0.94	89.87±0.86	87.18±2.64	92.76±1.85
Aco+Noi+MFCC	85.63±0.62	86.12±1.00	85.01±1.48	0.95	88.87±0.44	88.93±2.62	88.55±3.16

Table 4.9: Hypernasality detection using NHA, HAR and PHF features for /*u*/ vowel.

	Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**AUROC**	**Acc (%)**	**Sen (%)**	**Spe (%)**
NHA	71.43±2.55	61.80±7.86	79.02±2.67	0.75	74.05±3.81	63.67±3.81	82.16±3.81
HAR	76.08±1.11	79.86±1.47	73.17±1.10	0.82	74.05±3.81	63.67±10.24	82.16±2.24
PHF	80.49±1.51	80.18±3.21	80.66±4.14	0.87	81.89±1.12	80.11±4.54	83.17±2.95
NHA+HAR+PHF	82.65±0.59	85.38±2.15	80.49±1.39	0.90	84.25±1.18	86.40±3.45	82.59±2.06
VTC+PSR	80.35±1.38	84.83±3.35	76.62±0.48	0.89	84.44±2.14	88.00±5.97	81.60±1.46
MFCC	85.09±2.16	91.14±5.67	80.08±1.90	0.95	87.19±1.19	93.00±2.86	82.53±1.79
Aco+Noi+MFCC	83.35±2.01	89.79±4.32	78.11±1.38	0.92	85.41±2.20	90.49±5.42	81.33±1.70

for /u/ vowel than the other two features. Further, when individual NHA, HAR, and PHF features are combined as (NHA+HAR+PHF) feature, the performance increases compared to individual features for all three vowels. This shows that the NHA, HAR and PHF features are complementary. McNemars statistical test also shows that the increment in the performance for combined (NHA+HAR+PHF) feature compared to best performance among NHA, HAR and PHF features is statistically significant ($p < 0.001$). The performance is better for high vowels /i/ and /u/ compared to low vowel /a/ which supports the fact that the nasality is better perceived in high vowels [1]. The performance accuracy of 82.46%, 87.89% and 84.25% is obtained for the combined (NHA+HAR+PHF) feature at the phoneme level. The result shows that the best accuracy is obtained for /i/ vowel. The performance of combined (NHA+HAR+PHF) feature is better compared to VTC+PSR feature for /a/ and /i/ vowels and the performance is nearly equal for /u/ vowel. McNemars statistical test also shows that the increment in the performance for combined (NHA+HAR+PHF) feature compared to (VTC+PSR) feature is statistically significant ($p < 0.001$) for /a/ and /i/ vowels. However, the performance of the combined (NHA+HAR+PHF) feature is poor compared to the performance of MFCC and Aco+Noi+MFCC

features. The result shows that the sinusoidal model-based analysis of harmonics strength can capture the nasality evidence in a better way compared to the analysis of energy distribution, formant, noise, non-linearity or entropy in hypernasal vowels. But the performance of MFCC feature shows that the nasality evidence in hypernasal vowels is better captured by modeling the envelope of vowel spectrum rather than capturing the strength of harmonics present in the spectrum. Fig.4.11 (a)-(c) shows the ROC curve corresponding to NHA, HAR, PHF and combined (NHA+HAR+PHF) features for vowels /a/, /i/ and /u/ respectively. The curve is drawn for set 5 of the training-testing sets given in Table 3.6 their AUROC is given in Table 4.7, Table 4.8 and Table 4.9 respectively. The value of AUROC is high for the features having high performance.

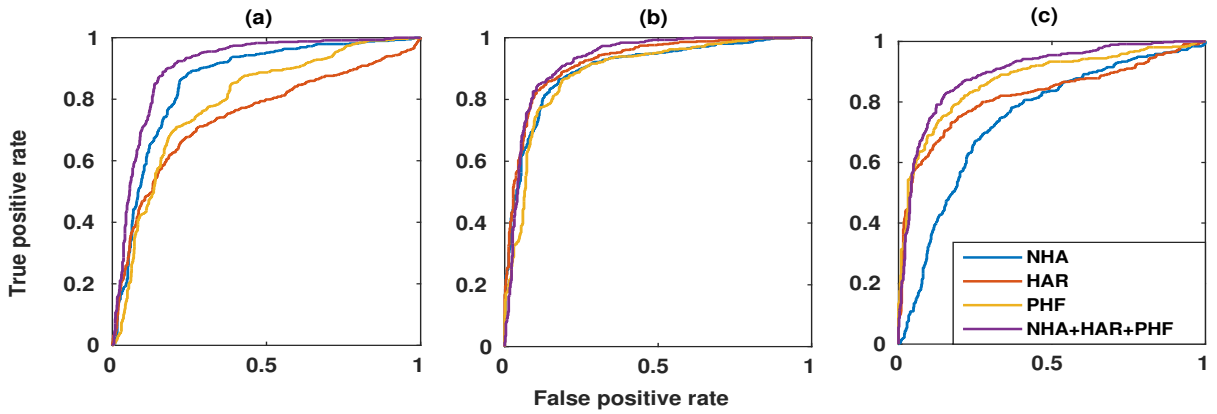

Figure 4.11: ROC curve for NHA, HAR, PHF and NHA+HAR+PHF features for different vowels. (a) for vowel /a/ (b) for vowel /i/ and (c) for vowel /u.

4.6 Summary

This chapter of the thesis proposes the NHA, HAR and PHF features for the classification of normal and hypernasal speech. The features are based on the strength of harmonics in the spectrum, and the harmonics strength is computed using the sinusoidal model of speech. The NHA feature is the magnitude of harmonics with respect to the maximum magnitude, HAR feature is the relative magnitude of a harmonics with respect to its previous harmonics magnitude, and the PHF feature is the frequency locations of prominent harmonics in the spectrum. The analysis showed that the nature of these features is different for hypernasal vowels compared to the normal vowels due to the presence of nasal formants and antiformants in the hypernasal vowels spectrum. The SD measures between feature dimensions and class labels are used to measure the discriminating capability of

each dimension of each feature, and the feature dimensions are arranged in decreasing order of SD measure. The higher SD values for the feature dimensions show the higher discriminating capability of that dimension, and it is shown using normalized histogram plots and mean and std values. The vowels /a/, /i/, and /u/ which are manually annotated from normal and hypernasal /papa/, /pipi/ and /pupu/ words are used as the database, and SVM classifier is used for the classification. Good classification performance for individual NHA, HAR and PHF features are obtained and it further increases for the combined (NHA+HAR+PHF) features. The (NHA+HAR+PHF) features perform better compared to VTC+PSR feature. However, the performance of (NHA+HAR+PHF) features is poor compared to MFCC and Aco+Noi+MFCC features. The MFCC feature captures the effect of the presence of nasal formant and antiformant in the hypernasal vowels by modeling the envelope of the spectrum. However, some literature shows that the MFCC feature has some issues for the high pitch children speech. so in the next chapter, some other cepstral features extracted from high-resolution spectrum, free from high pitch issues and spectral moment feature augmented with the cepstral features are explored for better performance of hypernasality detection.

5

Hypernasality detection using cepstral features

Publications

- **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Zero time windowing analysis of hypernasality in speech of cleft lip and palate children," in IEEE *Twenty Second National Conference (NCC)*, 1-6(2016).
- **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Hypernasality detection using Zero time windowing," in IEEE *International Conference on Signal Processing and Communications (SPCOM)*, 105-109(2018).
- **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat,"Pitch-Adaptive Front-end Feature for Hypernasality detection," in *Proc. Interspeech*, 372-376(2018).
- **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Detection and assessment of hypernasality in repaired cleft palate speech using vocal tract and residual features," *J. Acoust. Soc. Am.* 146(6), 4211-4223(2019).

Contents

Overview

The presence of nasal formant and antiformant pairs in the hypernasal vowels spectrum affects its resonance characteristics. Hence, the spectral envelope of hypernasal vowels deviates compared to the spectral envelope of normal vowels. Generally, the MFCC feature is used to model the spectral envelope. In this chapter, three cepstral features namely, Hilbert envelope of numerator of group delay function (HNGDF) feature, pitch-adaptive MFCC (PAMFCC) feature and spectral moment features augmented by low-order cepstral coefficients (SMAC) are used for hypernasality detection. These features capture the resonance characteristics of hypernasal vowels by modeling its spectral envelope. The HNGDF feature is the cepstral coefficients extracted from a high spectro-temporal Hilbert envelope of numerator of group delay (HNGD) spectrum. The HNGD spectrum can resolve the closely spaced nasal tract and oral tract resonances in the spectrum. The PAMFCC feature is extracted from the cepstrally smoothed spectrum instead of the magnitude spectrum. The smooth spectrum is free from the pitch harmonic effect present in low-frequency of magnitude spectrum where nasality evidence is also present. The SMAC feature is extracted from the band pass filtered signals of speech. The SMAC feature captures the information about the resonances and antiresonances in the spectrum along with the spectral envelope. The hypernasality detection accuracy of these features is compared with the accuracy obtained using the MFCC feature. The SVM classifier is used for hypernasality detection.

5.1 Introduction

The hypernasal vowels contain the nasal formant and antiformant pairs in their whole spectrum addition to their oral formants. The presence of nasal formant and antiformant pairs deviates the hypernasal vowels spectrum compared to normal vowels. In Chapter 4, the sinusoidal model-based features are used to captures the deviation in the strength of harmonics and it is found that the features are performing better compared to the temporal features explored in Chapter 3 for hypernasality detection. However, it is also found that the performance of sinusoidal model-based features is poor compared to the MFCC feature. The MFCC feature is the most widely used cepstral feature in the area of speech processing. It models the envelope of the spectrum which represents the vocal tract shape. However, the use MFCC feature for hypernasality detection in children speech may have some issues which affect its detection performance. This is because the MFCC feature is extracted from the magnitude spectrum of the signal which has limited frequency resolution and may not be sufficient

to resolve the closely spaced nasal and oral formants in the lower frequencies of the spectrum. Also, the Mel-filter bank employed on the magnitude spectrum to compute the MFCC feature is unable to sufficiently smooth out the pitch harmonic effect in the high pitch children speech. The insufficient smoothing causes the ripples in the modeled spectral envelope corresponding to the MFCC feature. This gives the high variance for the higher cepstral coefficients of the feature when computed across the speech database [45,89]. The variance may be higher in CP speech due to high pitch perturbation. The insufficient smoothing may also affect the ability of the MFCC feature in capturing the nasality evidence mainly present in lower frequencies around F_1 of the hypernasal vowels spectrum. Motivated from the aforementioned limitations of the MFCC feature, in this chapter some other cepstral features similar to MFCC or alternative to it are explored for better detection of hypernasality.

Three features namely, Hilbert envelope of numerator of group delay function (HNGDF) feature, pitch-adaptive MFCC (PAMFCC) feature and spectral moment features augmented by low-order cepstral coefficients (SMAC) are explored for the hypernasality detection. The HNGDF feature is cepstral coefficients computed from a high spectro-temporal Hilbert envelope of numerator of group delay (HNGD) spectrum [44]. The HNGD spectrum is derived from the phase spectrum rather than the magnitude spectrum as the group delay function is the negative derivative of the phase spectrum. The high temporal resolution of the HNGD spectrum is due to the zero time windowing (ZTW) operation in which the speech signal is multiplied with a highly decaying impulse-like window function of a short duration around a pitch period. The high spectral resolution of HNGD spectrum is due to an additive property of group delay function on the individual formants which is against the multiplicative nature of the magnitude spectrum. The PAMFCC feature is cepstral coefficients computed from the cepstral smoothed spectrum rather than the magnitude spectrum. A pitch adaptive liftering of cepstral coefficients with window size less than pitch period is done to compute the cepstral smoothed spectrum. The SMAC feature is the concatenation of spectral moments of signals obtained after passing the speech through a bank of band pass filters and low-order cepstral coefficients. The spectral moments capture the information about the formants and antiformants in the spectrum under the notion of the pyknogram. The low-order cepstral coefficients in the SMAC feature capture the energy and spectral envelope. The spectral moment feature has been proposed in the literature as an alternative to cepstral features [147]. The hypothesis about the HNGDF feature is that high temporal resolution will give the vocal tract characteristics of hypernasal speech within a pitch period. Also, the high

spectral resolution of the spectrum will resolve the closely spaced nasal and oral formants. Further, as the HNGD spectrum is derived from the phase spectrum rather than the magnitude spectrum, so the HNGDF feature may capture the complementary nasal evidence as captured by the MFCC feature for hypernasality detection. The hypothesis about the PAMFCC feature is that as it is computed from the cepstral smooth spectrum that eliminates the pitch harmonics effect present in the magnitude spectrum, so the PAMFCC feature can capture the nasality evidence present in the low-frequency band below around F_1 effectively. The hypothesis about the SMAC feature is that the feature captures the resonance as well as the spectral envelope characteristics of the hypernasal vowel spectrum. The hypernasality detection is done separately using HNGDF, PAMFCC and SMAC features and their performances are compared among themselves along with the performance obtained using the MFCC feature. The SVM classifier is used for hypernasality detection.

The rest of this chapter is organized as follows: Section 5.2 describes the process of computation of HNGDF feature. Section 5.3 describes the process of computation of PAMFCC feature. Section 5.4 describes the process of computation of SMAC feature. Section 5.5 describes the experimental setup and results and Section 5.6 summarizes the work.

5.2 HNGDF feature

Traditional methods of spectrum computation such as DFT spectrum, LP analysis, and cepstrum analysis uses a window size of 20-30 ms. Each of these methods has its limitations. The magnitude spectrum has pitch harmonic effect. In the LP spectrum, the number of peaks depends on the order of the LP model. The cepstrum based spectrum depends on the size of low time liftering window. Also, these methods give an average characteristic of the vocal tract system within the window segment. Hence these methods may not be suitable for capturing changes in highly non-stationary cases especially, within the closed and open glottis phase of one pitch period. To observe these changes, the analysis window should be small enough (around one pitch period). The traditional methods when used with small window size, give very poor frequency resolution [148]. The poor resolution spectrum cannot be used for hypernasality analysis because of its inability to resolve the closely spaced nasal and oral formants. Even the high spectral resolution spectrum such as group delay spectrum [149] and modified group delay spectrum [84] computed with small window size some times have spiky peaks that mask the peaks due to actual formants. So, a high spectro-temporal HNGD

spectrum is explored for hypernasality analysis and the feature is extracted from this spectrum for hypernasality detection.

The HNGD spectrum is computed using the operation called zero time windowing (ZTW) [44]. In this operation, the speech signal is multiplied with a highly decaying impulse-like window function of a short duration around a pitch period. The loss in spectral resolution due to ZTW operation is restored by successive differentiation in the frequency domain. This is because the windowing in the time domain is an approximation to integration operation in the frequency domain. The spectral resolution is further improved by the use of the group delay function. The ZTW operation can be performed at any sampling instant. So, the HNGD spectrum is an instantaneous spectrum that can give the vocal tract system characteristics at each sampling instant. The HNGDF feature is derived from the HNGD spectrum after computing the discrete cosine transform (DCT) of the logarithmic HNGD spectrum. This section of the chapter discusses the ZTW operation, steps to compute the HNGD spectrum, analysis of hypernasal speech using the HNGD spectrum and method to compute the HNGDF feature from the HNGD spectrum.

5.2.1 Zero-time windowing of speech

The frequency response of zero frequency resonator (ZFR) [123] is

$$|H(\omega)| = \left| \frac{1}{(1-z^{-1})^2} \right|_{z=e^{j\omega}} = \frac{1}{2(1-\cos\omega)} = \frac{1}{4\sin^2\omega/2} \tag{5.1}$$

The filtering of speech signal through a ZFR is represented by following operation in time domain

$$y[n] = 2y[n-1] - y[n-2] + s[n] \tag{5.2}$$

where, $s[n]$ is DC bias removed speech input and $y[n]$ is the output of the resonator. The frequency domain equivalence of above operation is to multiply the spectrum of speech $s[n]$ with a window function given by the frequency response of resonator $|H(\omega)|$. This is equivalent to integrating the speech signal twice in time domain to provide a sharp role off.

ZTW is a time domain operation analogous to ZFF operation in frequency domain. The time domain signal is multiplied with a window function similar in the shape of frequency response window

of the ZFR. The window is called zero time window function and is given by

$$w_1[n] = \begin{cases} 0, & n = 0 \\ \frac{1}{(4\sin^2(\pi n/(2N)))}, & n = 1, 2, ..., N-1 \end{cases} \tag{5.3}$$

where, N is the window length. This window function or analogous window in ZFR can be approximated to $\frac{1}{n^2}$ for smaller value of n and $N >> M$, which is equivalent to integrating the signal spectrum in frequency domain. M is the length of the speech segment and N is chosen to be the DFT length. Window function gives more weight to the zero time samples and less weight to all other samples of the segmented signal.

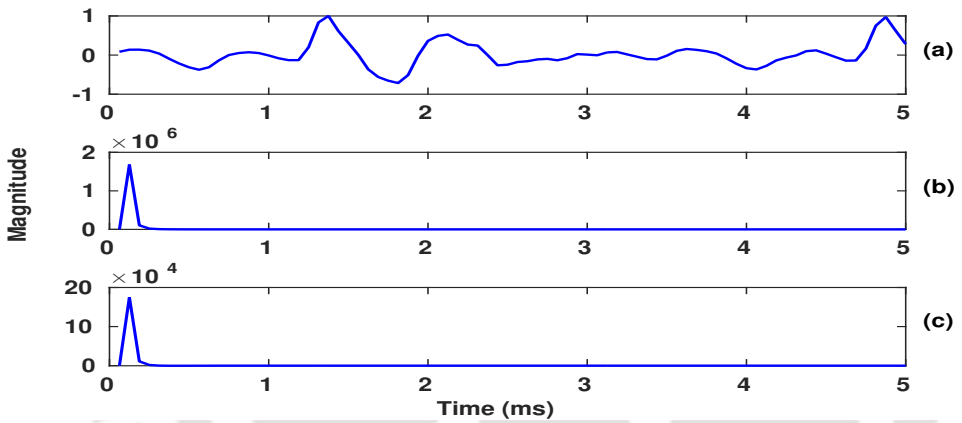

Figure 5.1: Illustration of ZTW of speech signal.(a) Short segment of speech waveform. (b) ZTW function $w_1(n)$. (c) ZT windowed speech waveform $x(n)$.

The concept of ZTW is shown in Fig. 5.1. The speech segment of length 5 ms is shown in Fig. 5.1 (a). The ZTW function is shown in Fig. 5.1 (b). The multiplication of the speech signal and ZTW function is shown in Fig. 5.1 (c). The nature of the windowed signal looks almost like an impulse since the signal has a high value of amplitude only around zero time $n = 0$ (reference time) and hence the name zero time windowing. Since the spectrum of the windowed signal will show the characteristics of the signal around $n = 0$, such characteristics are also called instantaneous spectral characteristics.

5.2.2 HNGD spectrum

The HNGD spectrum has the higher temporal and spectral resolution. It is an instantaneous spectrum. Hence it can be used to detect the nasal and oral formants in hypernasal vowels at each sampling instant. It is based on ZTW method where a segment of speech. The various steps involved in extracting the HNGD spectrum are as follows:

- Consider the differenced speech signal $s[n]$ at a sampling frequency of f_s Hz. The signal is differenced to remove low frequency bias in the signal.
- Consider $s[n]$ of M samples starting from an arbitrary reference sampling instant set at $n = 0$ to end at $M - 1$.
- Take DFT length $N >> M$. Zero pad the signal $s(n)$ to make it of length $N >> M$.
- Multiply N length $s[n]$ segment with a window function $w_1^2[n]$. This window function emphasizes the values near the $n = 0$ sampling instant.
- The truncation effect at the end of window ($M - 1$ sampling instant in time domain) results ripple in the frequency domain. This effect is reduced by using another window.

$$w_2[n] = 2(1 + \cos(\pi n/M)) = 4\cos^2(\pi n/2M), \quad n = 0, 1, ..., M - 1.$$

- Compute N-point DFT of the double windowed signal, i.e. of $x[n] = w_1^2[n]w_2[n]s[n]$. The square magnitude spectrum of windowed signal is smooth due to equivalent four time integration in frequency domain.
- To highlight the spectral characteristics, the numerator of the group delay (NGD) function $g[k]$ of the windowed signal $x[n]$ is computed [150]. The NGD function $g[k]$ is the numerator part of the group delay function proposed in [148]. The group delay function for the signal $x[n]$ is given by [151],

$$\tau[k] = \frac{X_R[k]Y_R[k] + X_I[k]Y_I[k]}{X_R^2[k] + X_I^2[k]}, \quad k = 0, 1, ..., N - 1. \tag{5.4}$$

where $X_R[k]$ and $X_I[k]$ are the real and imaginary parts of the N-point DFT $X[k]$ of $x[n]$ respectively and $Y_R[k]$ and $Y_I[k]$ are the real and imaginary parts of the N-point DFT $Y[k]$ of $y[n] = nx[n]$ respectively. The numerator of the group delay function $g[k]$ is given by

$$g[k] = X_R[k]Y_R[k] + X_I[k]Y_I[k], k = 0, 1, ..., N - 1. \tag{5.5}$$

The group delay function has high frequency resolution due to additive property ($\tau[k] \propto |H[k]|^2$ around the formants frequency) [152]. The spectral resolution further increases in NGD due to ignorance of denominator term ($g[k] \propto |H[k]|^4$ around the formants frequency) [148].

- To further highlight the spectral characteristics like peaks corresponding to formants, NGD

function is differenced twice (DNGD) in the frequency domain. This operation is opposite to twice integration operation due to the multiplication of speech signal by the ZTW function.

- The low amplitude spectral peaks in the double differenced NGD spectrum gets suppressed due to valleys. The HE of the double differenced NGD spectrum is computed to remove the effect of spectral valleys. The resulting envelope is called the HNGD spectrum.

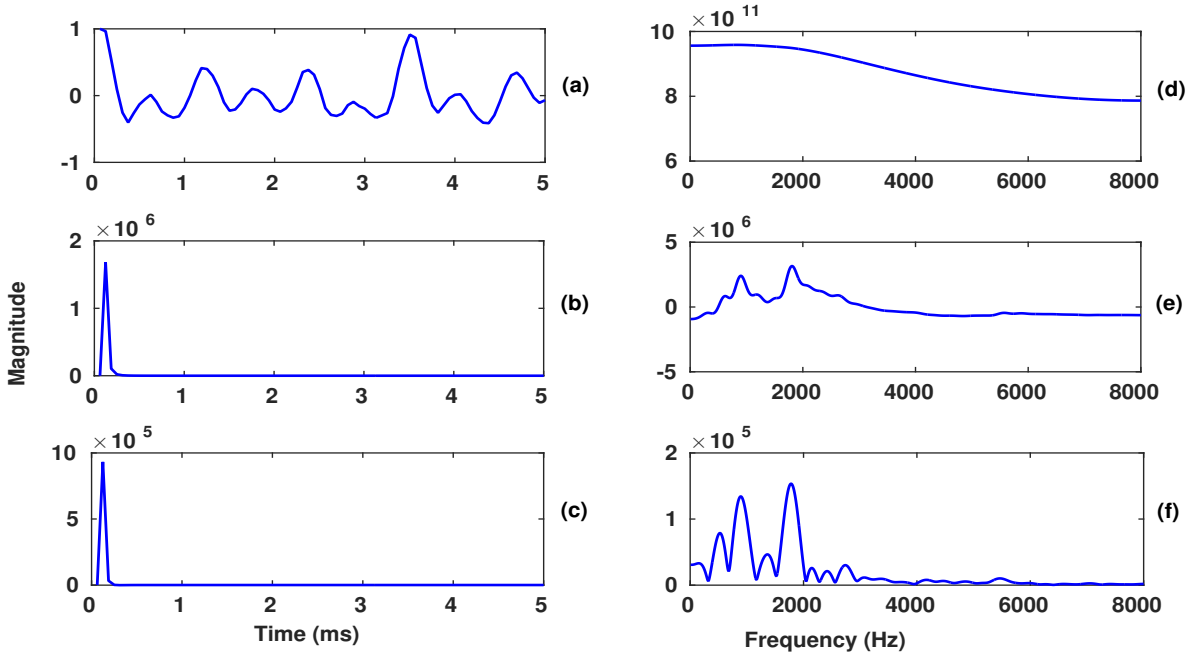

Figure 5.2: Illustration of ZTW method for hypernasality detection. (a) Short segment (5 ms) of hypernasal vowel /a/ and ZTW function. (b) Combined window function $w(n) = w_1^2(n)w_2(n)$. (c) Windowed speech waveform $x(n) = s(n)w(n)$. (d) NGD spectrum of (c). (e) Double derivative of (d). (f) HE of (e)

Fig. 5.2 shows the complete illustration steps to compute HNGD spectrum based on ZTW method. Fig. 5.2 (a) shows the 5 ms speech waveform and ZTW function. The highly exponential nature of the ZTW function can be observed. It gives high weight to the samples near zero time of the speech segment than the other samples in the segment. Multiplication with the ZTW function in time domain is equivalent to integration of the speech spectrum in frequency domain. The combined window function $w(n) = w_1^2(n)w_2(n)$ is shown in Fig. 5.2 (b). The multiplication of combined window function with the speech signal in Fig. 5.2 (a) is shown in Fig. 5.2 (c). The NGD spectrum of Fig. 5.2 (c) is shown in Fig. 5.2 (d). The twice successive derivative of Fig. 5.2 (d) is shown in Fig. 5.2 (e). The Hilbert envelope of the Fig. 5.2 (e) is plotted in Fig. 5.2 (f) which is the final HNGD spectrum. The dominant peaks in the HNGD spectrum are considered as formants. Fig. 5.2 (f) shows the HNGD

spectrum for hypernasal /a/ vowel. It can be observed that there are two dominant peaks below 1000 Hz. One below 500 Hz and another around 700 Hz. The peak around 700 Hz is the F_1 of /a/ vowel and the peak below 500 Hz is the nasal formant. This shows that the HNGD spectrum is capable of resolving the closely spaced nasal and oral formants in hypernasal vowels.

Another important advantage of the HNGD spectrum is that it can be computed at each sampling instant of speech. This is because the ZTW operation can be applied at each sampling instant. As the GCIs are a high signal to noise ratio (SNR) instants, the HNGD spectrum at these instants is more prominent than the other instants. Fig. 5.3 shows the HNGD spectrum at each sampling instant of a speech signal having three-pitch periods. In each period, at one sampling instant, the HNGD spectrum is more prominent compared to others. These instants are GCIs. The computation of the HNGD spectrum at the GCIs of hypernasal speech may capture better nasality evidence compared to other instants.

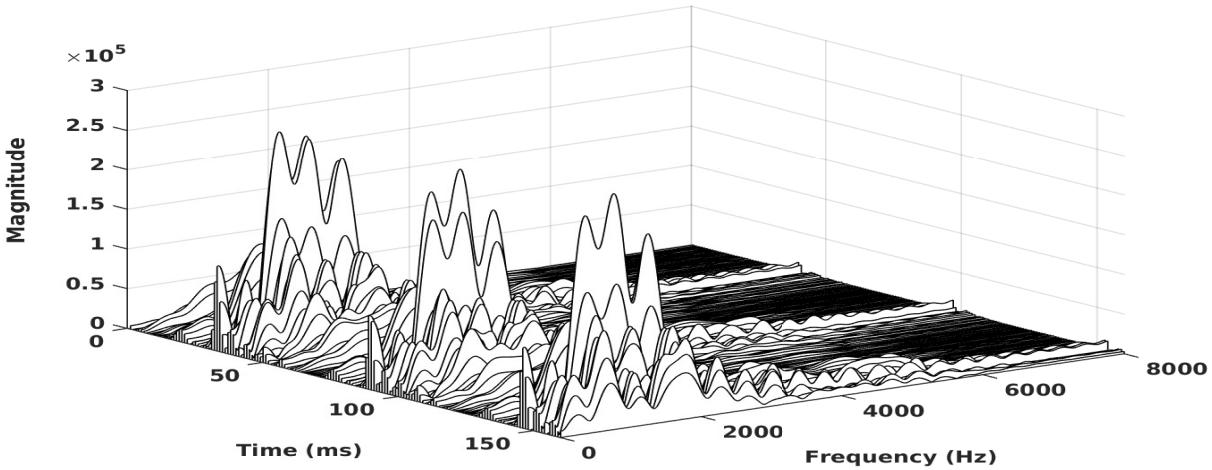

Figure 5.3: HNGD spectrum at each sampling instant within three pitch period

5.2.3 Analysis of hypernasal speech using HNGD spectrum

The spectral characteristics of hypernasal vowels differ from the normal vowel characteristics due to the presence of nasal formant in the vicinity of F_1. To show the usefulness of the HNGD spectrum for resolving the closely spaced nasal and oral formants in the hypernasal vowels spectrum, the HNGD spectrum of normal and hypernasal vowels are analyzed. The HNGD spectrum gives the vocal tract characteristics of hypernasal speech within a pitch period due to the ZTW operation. Further, the instantaneous nature of the spectrum can be used for finding the vocal tract characteristics at the

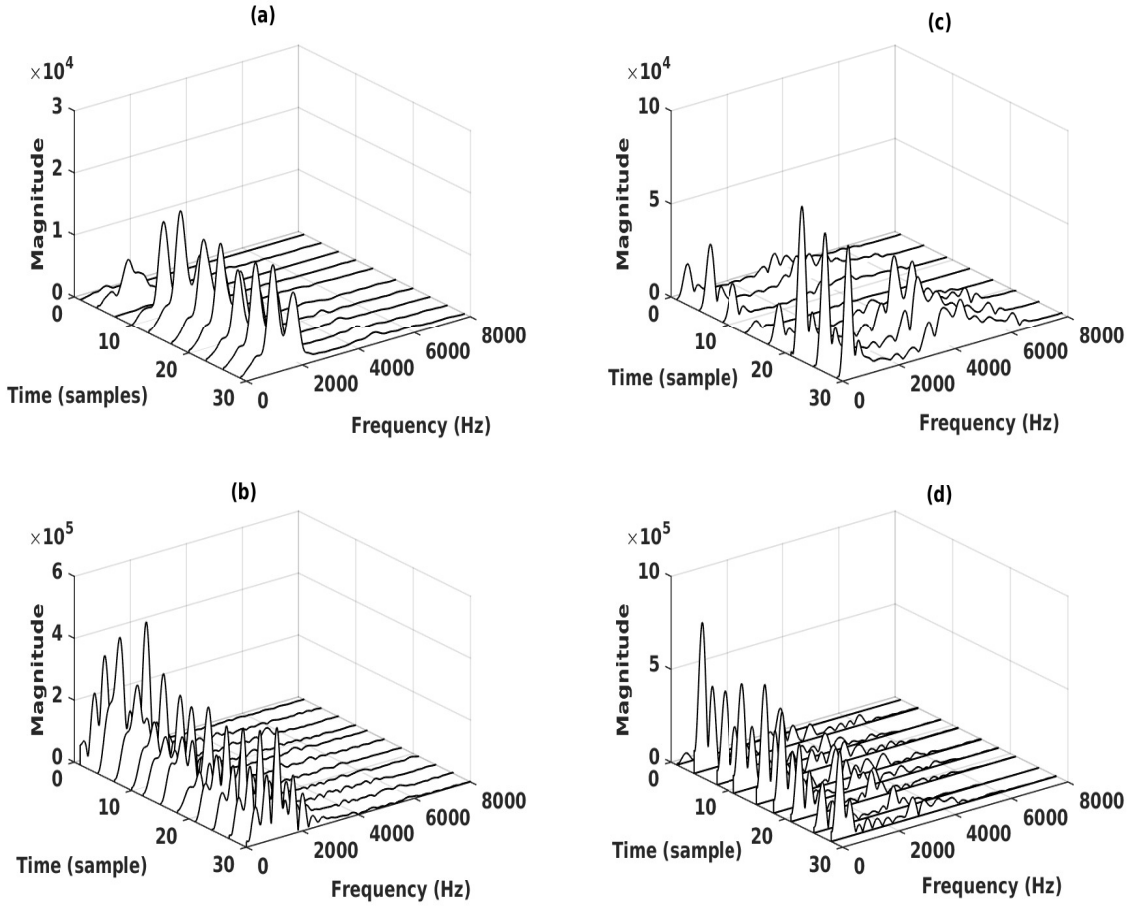

Figure 5.4: HNGD spectrum of vowel sounds. (a) Normal vowel /a/ with first formant around 700 Hz. (b) Hypernasal vowel /a/ with first formant around 700 Hz and resolved additional nasal peak below 500 Hz. (c) Normal vowel /i/ with first formant around 400 Hz. (d) Hypernasal vowel /i/ with first formant around 400 Hz and resolved additional nasal peak around 1000 Hz.

epoch locations. Fig. 5.4 (a)-(d) shows the HNGD spectrum at epoch locations for the normal and hypernasal vowel sounds. The epoch locations are determined using the method proposed in [123]. Fig. 5.4 (a)-(b) represents the HNGD spectrum of normal and hypernasal /a/ vowel, respectively, and the Fig. 5.4 (c)-(d) represents the HNGD spectrum for the normal and hypernasal /i/ vowel, respectively. It can be observed from Fig. 5.4 (b) that the HNGD spectrum is capable of resolving the additional nasal peak below 500 Hz and the F_1 around 700 Hz for hypernasal /a/ vowel. Similarly, it can be observed from Fig. 5.4 (d) that the HNGD spectrum is capable of resolving the F_1 of vowel /i/ around 400 Hz and the addition nasal peak around 1000 Hz. So, the analysis shows that the HNGD spectrum can capture the nasality cues present in hypernasal vowels by resolving the nasal and oral

formants.

5.2.4 Computation of HNGDF Feature from HNGD spectrum

To parameterize the HNGD spectrum into a feature for hypernasality detection, the most commonly used approach called Homomorphic processing of the HNGD spectrum is done. The block diagram of the approach is shown in Fig. 5.5. The logarithmic HNGD spectra are derived by taking the logarithm of HNGD spectra computed at the epoch locations. The DCT (II) [153] of the logarithmic HNGD spectra gives the proposed cepstral feature termed as HNGD function (HNGDF) feature.

Figure 5.5: Block diagram showing the steps of conversion of HNGD spectra to the HNGD function (HNGDF) feature

5.3 PAMFCC feature

The MFCC feature is extracted from the DFT magnitude spectrum which has the pitch harmonic effect. The high pitch value, as in the case of children speech, affects the statistical characteristics of higher cepstral coefficients of the MFCC feature. The higher coefficients show high variance because Mel-filterbank of the MFCC feature is unable to smooth out the pitch harmonic effect. This may affect the performance of the MFCC feature when the feature is used for hypernasality detection. To overcome this limitation of the MFCC feature, the PAMFCC feature is explored for hypernasality detection. The PAMFCC feature is computed from the cepstral smoothed spectrum which is obtained after pitch adaptive liftering of cepstral coefficients derived from the magnitude spectrum. The pitch adaptive liftering eliminates the pitch harmonic. The Mel-filter bank is employed on the cepstral smoothed spectrum to compute the PAMFCC feature. The elimination of pitch harmonic effect from the low-frequency region of the spectrum where the nasality evidence is present in hypernasal vowels helps the PAMFCC feature to capture the nasality evidence in a better way. Also, the variance of higher coefficients reduces. These factors may lead to better hypernasality detection performance. This section of the chapter, discusses the effects of pitch on MFCC feature, steps to compute PAMFCC feature and significance of PAMFCC feature for hypernasality detection.

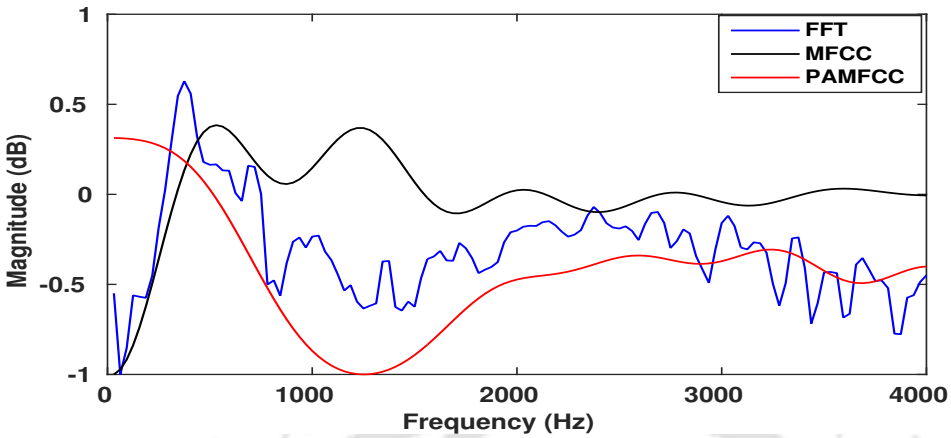

Figure 5.6: Plots of smooth spectra corresponding to MFCC and PAMFCC feature along with STFT magnitude spectrum for the vowel /i/ of high pitch CP children speech.

5.3.1 Effect of pitch on MFCC feature

For the extraction of the MFCC feature from the speech signal, firstly the framing of the signal using overlapping Hamming/Hanning windows is done and then the magnitude spectrum of each frame is computed using the short-time Fourier transform. Next, the Mel-scale warping of the magnitude spectrum is done using triangular filters having nonuniform bandwidth. The DCT of the log-energies obtained as an output of the Mel-filterbank gives the MFCC feature. The MFCC feature model the magnitude spectrum and gives the smooth spectral envelope representing the vocal tract characteristics. Hence it is expected that the feature is free from the pitch harmonics effect present in the magnitude spectrum. But the studies [45, 89] show that for high pitch signals like children speech, the Mel-filter bank employed in the MFCC feature extraction is unable to sufficiently smooth out the pitch harmonics present in the magnitude spectrum. Hence the MFCC feature gets affected for the high pitch signal. The smoothed spectral envelope corresponding to the affected MFCC feature contains ripples in the lower frequency region. The ripples give a high variance for the higher coefficients of the MFCC feature.

Fig. 5.6 shows the DFT magnitude spectrum and the smoothed spectra corresponding to the MFCC feature for vowel /i/ of CP children speech. The smooth spectra corresponding to the MFCC feature is derived by taking the 128-point inverse discrete cosine transform of 13-dimensional MFCC feature. Ripples in smoothed spectra corresponding to MFCC feature in the lower frequency region can be observed in Fig. 5.6. The effect of ripples on the variance of the 13-dimensional MFCC feature

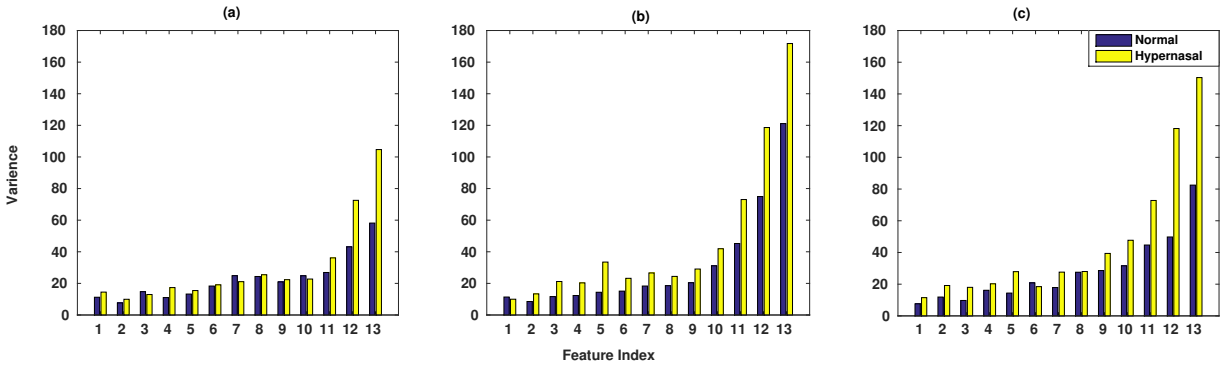

Figure 5.7: Plot showing the variance (in bar) for each coefficients of 13-dimensional MFCC feature extracted for normal and hypernasal vowels from entire database. (a) for vowel /a/, (b) for vowel /i/ and (c) for vowel /u/ vowels

for normal and hypernasal /a/, /i/ and /u/ vowels is shown in Fig. 5.7 (a)-(c) in the form of bar plot. It can be observed from the bar plot that the variance is higher for higher coefficients (11-13 coefficients) of the MFCC feature. Further, it can also be observed that the variance is more for the hypernasal vowels compared to the normal vowels. This is because there is high pitch perturbation in CP speech due to the lack of control of vocal fold vibration [35]. The high pitch perturbation in CP speech can be ascertained by measuring the mean and standard deviation of the pitch for all three vowels /a/, /i/ and /u/ from the entire database discussed in Section 3.2. Table 5.1 shows the mean and std of the pitch for normal and hypernasal vowels from the entire database. The pitch is measured using the event-based instantaneous fundamental frequency method proposed in [143]. It can be observed that the mean, as well as std of the pitch, is high for hypernasal vowels compared to normal vowels. Also, std which shows the pitch variation is high for normal and hypernasal vowels, but it is higher for hypernasal vowels compared to the normal vowels. The ripples in low-frequency smooth spectra corresponding to the MFCC feature and high variance in higher coefficients of the MFCC feature may affect the classification accuracy of normal and hypernasal vowels.

Table 5.1: Mean and standard deviation (std) of pitch in Hz computed from entire database having normal and hypernasal vowels.

Vowel	Normal		Hypernasal	
	mean	std	mean	std
/a/	282.57	62.10	299.89	63.05
/i/	299.26	57.18	315.42	69.30
/u/	279.94	48.18	311.47	95.04

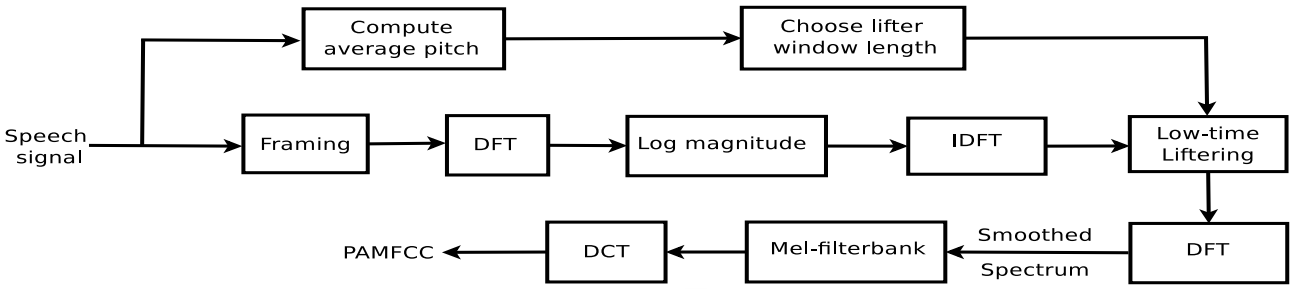

Figure 5.8: Block diagram for the extraction of the pitch-adaptive MFCC feature by applying adaptive-liftering for spectral smoothening.

5.3.2 Computation of PAMFCC feature

The PAMFCC feature is originally proposed for robust children's automatic speech recognition in [154]. The feature is free from the pitch harmonics effect and also deals with the high pitch perturbation in CP speech. This is because a pitch adaptive low time liftering of cepstral coefficients derived from the magnitude spectrum is done while computing the feature. The block diagram for the extraction of PAMFCC feature is shown in Fig. 5.8. The procedure for deriving the PAMFCC feature are as follows:

- Compute the log magnitude spectrum of each frame of the speech signal using the short-time Fourier transform (STFT) with a fixed duration hamming window.
- Obtain the cepstral representation through the inverse discrete Fourier transform (IDFT) of the magnitude spectrum.
- Apply a pitch adaptive low time liftering on the cepstral representation because it retains the periodicity of the speech excitation. The pitch adaptive liftering smooth the pitch harmonic. Take the duration of low-time lifter $L = \frac{F_s}{F}$, where F_s is the sampling frequency and F is the average pitch value for the whole utterance. In this work, the pitch of the utterance is detected using the zero frequency filtered signal as proposed in [143].
- Take the DFT of the liftered cepstrum to obtain the smoothed cepstral spectrum.
- Employ Mel-filter bank on the smoothed cepstral spectrum and compute the log-energies for each filter.
- Take the DCT of the log-energies to find the cepstral coefficients.

- The lower coefficients are PAMFCC feature.

Fig. 5.6 shows the smoothed spectra corresponding to the PAMFCC feature. It can be observed from Fig. 5.6 that the low-frequency ripples are smoothed out in the PAMFCC features. Hence, the feature can capture the low-frequency nasality evidence in a better way. The smoothing pitch also results in the reduction of variance for higher coefficients in the PAMFCC feature. The detection of hypernasality using the PAMFCC feature may give better classification accuracy of normal and hypernasal vowels.

5.4 SMAC feature

The SMAC feature is originally proposed for robust automatic speech recognition in [155]. The feature is computed by first passing the speech signal through a bank of band pass filters and then computing the spectral moment of each band pass filtered signals. The spectral moments capture the information about the formants and antiformants in the spectrum under the notion of the pyknogram [156]. However, the pyknogram solely does not model the relative importance of each formants and antiformants in the spectrum. Hence, the spectral moments are augmented with the low-order cepstral coefficients to forms the SMAC feature. The low-order cepstral coefficients in the SMAC feature capture the energy and spectral envelope. The presence of both nasal formants and antiformants deviates the resonance as well as the spectral envelope characteristics of the hypernasal vowels spectrum. Hence, the values of the SMAC feature may differ for normal and hypernasal speech.

5.4.1 Computation of SMAC feature

SMAC feature captures the spectral deviation present in the entire spectrum of hypernasal vowels compared to normal vowels. The deviation is mostly around the formants and antiformants in the spectrum, and it affects the spectral envelope. SMAC feature is based on the first normalized central spectral moment augmented by few low-order cepstral coefficients. The SMAC feature is computed as follows [155].

- Filter the speech frame $s(n,t)$, $n = 0, 1, ..., N-1$ at a given time instant t using a bank of K bandpass filters. Let the output of k^{th} filter be expressed in spectral domain as,

$$S_k(\omega, t) = S(\omega, t)H_k(\omega), \tag{5.6}$$

where $S(\omega, t)$ is the spectra of t^{th} frame and $H_k(\omega)$ is the frequency response of the k^{th} bandpass filter where $k = 1, ..., K$.

- For each band, compute the m^{th} central power spectral moment defined as,

$$M^m(k,t) = \int_0^\pi |S_k(\omega,t)|^2 (\omega - \omega_k)^m d\omega, \qquad m = 0, 1, ... \tag{5.7}$$

- Compute the normalized central moment as

$$N^m(k,t) = \frac{M^m(k,t)}{M^0(k,t)}. \tag{5.8}$$

The K normalized first central moments, denoted by $\{N^1(1,t), ..., N^1(K,t)\}$, appended with first two cepstral coefficients form the SMAC feature. The cepstral coefficients capture the spectral slope.

- Use Mel-spaced Gabor filterbank having frequency response for the kth real Gabor filter as

$$H_k(\omega) = \frac{\sqrt{\pi}}{2\beta} \exp\left\{-\frac{(\omega-\omega_k)^2}{4\beta^2}\right\}, \quad \omega > 0 \tag{5.9}$$

where the bandwidth of filters is controlled by the parameter β to extract the SMAC feature.

Fig. 5.9 (a)-(b), respectively, show the DFT spectrum of normal and hypernasal /i/ vowel superimposed with their first central spectral moment estimation for Mel-spaced Gabor filterbanks. The spectral moment estimates are shown with the vertical line, and the filterbank central frequencies are shown with triangles. It can be observed from Fig. 5.9 that the first central spectral moment of each filterbank is different for normal and hypernasal vowels. This suggests that the spectral moments are capable of capturing the different resonance characteristics in normal and hypernasal vowels. The lower cepstral coefficients capture the spectral slope and when these coefficients are augmented with the spectral moments to form SMAC feature, the feature captures the resonance as well as the slope of the spectrum. So SMAC feature can be used for hypernasality detection.

5.5 Experiment and results

In this section, experimental setup and result of normal vs. hypernasal vowel classification using the HNGDF, PAMFCC, and SMAC features are presented. The performance of these features is

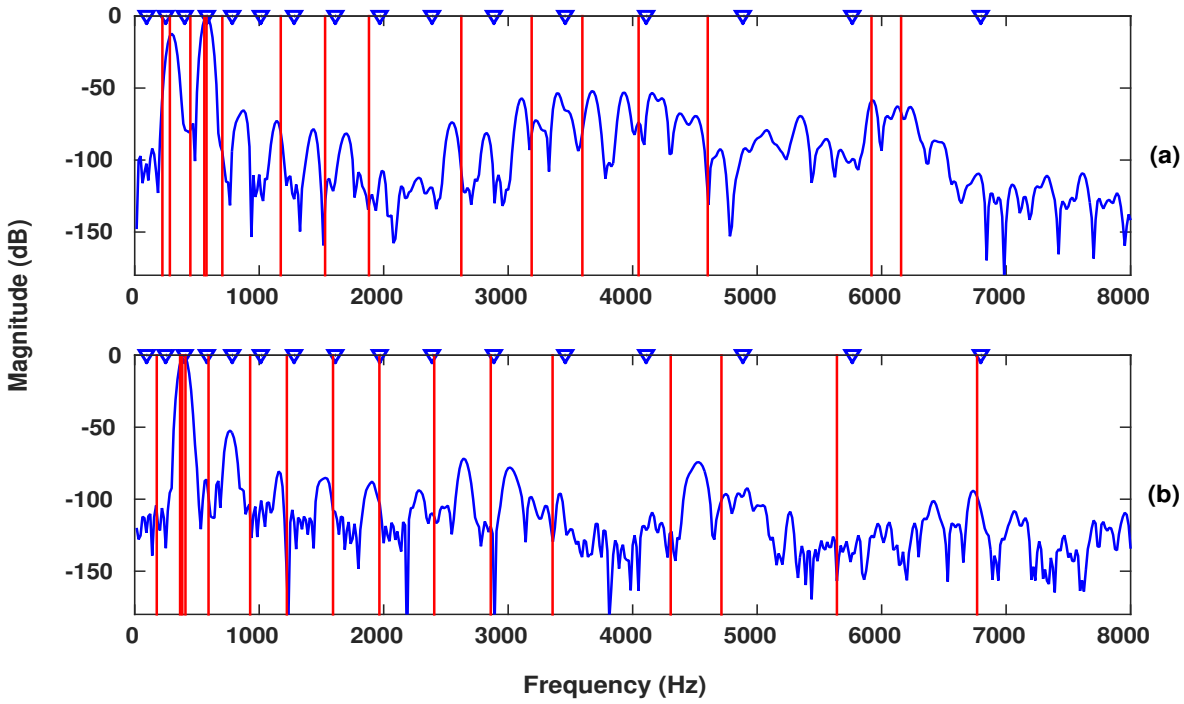

Figure 5.9: DFT spectrum of /i/ vowel superimposed with the first central spectral moment estimations for Mel-spaced Gabor filterbanks. (a) for normal /i/ vowel and (b) for hypernasal /i/ vowel. The spectral moment estimates are shown with vertical line and the filterbank central frequencies are shown with triangles.

compared with the performance of the MFCC feature as the MFCC feature is performing best among all baseline features. The performance is evaluated using the SVM classifiers.

5.5.1 Experimental setup

The HNGDF feature is extracted from the HNGD spectrum computed at every epoch locations of speech signal, which is further processed framewise by sampling it at the framing instant. To compute HNGDF feature, HNGD spectrum is modelled by 13 DCT coefficients. The PAMFCC and SMAC features are directly computed framewise with the frame size of 20 ms and frameshift of 10 ms. A total of 30 filterbanks are used to compute 30 cepstral coefficients and only lower 13 coefficients ($C_1 - C_{13}$) are kept as a 13-dimensional PAMFCC feature. The values of β, number of Gabor filter (K), and filter bandwidth taken for the computation of SMAC feature are 2, 16, and 236 Mels, respectively as suggested in [155]. The C_0 coefficient is ignored for the computation of SMAC features. Only C_1 coefficient is augmented with the first central spectral moment values to compute the SMAC feature. Hence SMAC is a 17-dimensional feature. Five classification experiments using five different training-

testing sets are performed for each vowel. The database, five different training-testing sets, and the RBF kernel parameters (c, γ) are used the same that is used in Chapter 3.

5.5.2 Results

Like in previous chapters, the classification results of normal vs. hypernasal vowels are presented at the frame as well as the phoneme level. The results are presented in terms of accuracy (Acc), sensitivity (Sen) and specificity (Spe) parameters. The results are presented in the form of mean and std of results obtained for five training-testing sets for each feature. Table 5.2, Table 5.3 and Table 5.4 show the classification performance of MFCC, HNGDF, (HNGDF+MFCC), PAMFCC and SMAC features for /a/, /i/ and /u/ vowels, respectively. It can be observed that the performance of the HNGDF feature is poor compared to the MFCC feature for /a/, /i/ and /u/ vowels. The performance difference between HNGDF and MFCC features is low for /a/ and /i/ vowels but it is high for /u/ vowel. The reason for the poor performance of the HNGDF feature may be the nature of the HNGD spectrum. The HNGD spectrum is a smooth envelope of magnitude spectrum which may lose some information due to smoothing. The comparatively more poor performance of the HNGDF feature for /u/ vowel maybe because of the presence of oral formants F_1, F_2 and nasal formant all below 1000 Hz and the smoothing of the spectrum. The HNGD spectrum may not be able to resolve three formants below 1000 Hz in /u/ vowel. But when the HNGDF feature is concatenated with the MFCC feature, the performance of 26-dimensional (HNGDF+MFCC) feature is better compared to MFCC feature for /a/ and /i/ vowels, and it is worse for /u/ vowel. The increment in performance for /a/ and /i/ vowels shows that the HNGDF feature and MFCC feature are complementary to each other because the former feature is extracted from the phase spectrum and the latter is extracted from the magnitude spectrum of the speech signal. McNemars statistical test also shows that the increment in the performance for combined (HNGDF+MFCC) feature compared to the MFCC for /a/ and /i/ vowels is statistically significant ($p < 0.001$).

The results also show that the performance for PAMFCC and SMAC features is better compared to MFCC feature for /a/, /i/ and /u/ vowels. The better performance of PAMFCC feature can be attributed to the smoothing out of the pitch harmonic effect present in the DFT magnitude spectrum of the signal. The reason for the better performance of the SMAC feature is due to the capturing of nasal formant and antiformant information present in the hypernasal speech spectrum along with the spectral envelope in a better way compared to MFCC feature. The PAMFCC feature is giving the

best performance of 89.56% and 92.30% for /a/ and /i/ vowels respectively. The SMAC feature is giving the best performance of 89.19% for /u/ vowel. In this case also, the McNemars statistical test shows that the increment in the performance for PAMFCC and SMAC features compared to MFCC feature is statistically significant ($p < 0.001$).

Table 5.2: Hypernasality detection using HNGDF, PAMFCC and SMAC features for /*a*/ vowel.

	Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**AUROC**	**Acc (%)**	**Sen (%)**	**Spe (%)**
MFCC	78.79±1.53	79.61±3.36	78.25±2.98	0.87	83.51±1.97	80.58±4.31	85.55±3.83
HNGDF	76.91±3.86	69.48±4.21	81.96±4.12	0.86	83.14±4.20	75.09±4.43	89.20±5.64
HNGDF+MFCC	82.20±1.56	82.24±4.92	82.17±1.13	0.91	87.76±2.50	85.20±6.48	89.56±1.27
PAMFCC	85.23±1.15	84.58±4.59	85.71±3.55	0.93	89.56±1.37	89.88±6.19	89.39±4.54
SMAC	80.08±1.16	78.38±3.60	81.24±1.78	0.89	84.46±2.07	81.15±5.88	86.78±2.37

Table 5.3: Hypernasality detection using HNGDF, PAMFCC and SMAC features for /*i*/ vowel.

	Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**AUROC**	**Acc (%)**	**Sen (%)**	**Spe (%)**
MFCC	84.68±0.56	83.74±1.98	85.29±1.27	0.94	89.87±0.86	87.18±2.64	92.76±1.85
HNGDF	80.95±0.47	77.60±3.44	84.66±4.68	0.89	86.86±1.44	81.55±2.68	92.83±3.15
HNGDF+MFCC	86.88±1.02	85.37±2.08	88.46±0.88	0.95	92.03±3.13	88.94±8.00	95.21±1.74
PAMFCC	87.07±2.47	85.16±7.71	88.80±2.58	0.95	92.30±1.31	88.93±2.87	95.91±0.66
SMAC	85.83±0.67	81.51±2.64	90.38±2.67	0.93	89.91±1.99	84.77±3.26	94.58±3.69

Table 5.4: Hypernasality detection using HNGDF, PAMFCC and SMAC features for /*u*/ vowel.

	Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**AUROC**	**Acc (%)**	**Sen (%)**	**Spe (%)**
MFCC	85.09±2.16	91.14±5.67	80.08±1.90	0.95	87.19±1.19	93.19±2.86	82.53±1.79
HNGDF	69.62±1.99	61.99±1.95	75.87±3.42	0.69	72.66±3.42	62.32±3.71	80.93±5.90
HNGDF+MFCC	84.49±1.25	90.40±4.59	79.62±2.55	0.94	86.15±1.65	91.49±6.06	81.87± 2.47
PAMFCC	86.33±2.30	85.91±6.25	86.67±1.12	0.95	87.55±2.47	84.43±6.37	90.07± 0.79
SMAC	86.07±0.81	90.24±3.84	82.62±2.15	0.95	89.19±1.04	92.99±3.02	86.13± 3.72

Fig. 5.10 (a)-(c) shows the ROC curve corresponding to MFCC, HNGDF, HNGDF+MFCC, PAMFCC and SMAC features for vowels /a/, /i/ and /u/ respectively. The curve is drawn for set 5 of the training-testing sets given in Table 3.6 and their AUROC is given in Table 5.2, Table 5.3 and Table 5.4 respectively. The value of AUROC is high for the features having high performance.

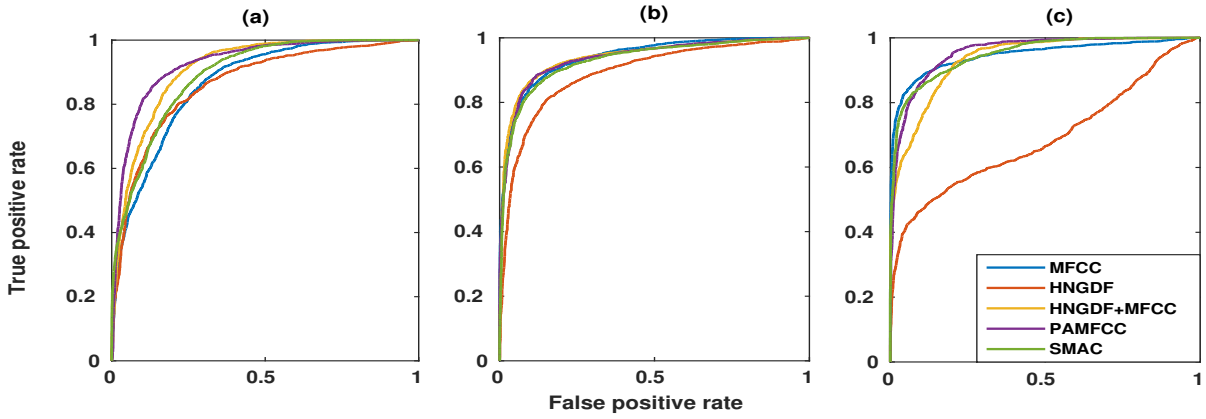

Figure 5.10: ROC curve for MFCC, HNGDF, HNGFDF+MFCC, PAMFCC and SMAC features for different vowels. (a) for vowel /a/ (b) for vowel /i/ and (c) for vowel /u.

5.6 Comparison of performance using different features and their combinations

In this section, the performance of hypernasality detection using the individual temporal, sinusoidal model-based, cepstral features, and their inter combination are compared. The inter combination of features is done at the score level. Table 5.5, Table 5.6 and Table 5.7 show the performance of different features and their inter combination for /a/, /i/ and /u/ vowels respectively. Each table shows the performance for temporal feature (VTC+PSR), sinusoidal model-based feature (NHA+HAR+PHR), cepstral features PAMFCC, SMAC and baseline MFCC feature. The tables are also showing the performance for combination of temporal feature with the sinusoidal model-based feature (VTC+PSR+NHA+HAR+PHF), combination of temporal feature with cepstral features (VTC+PSR+PAMFCC) and (VTC+PSR+SMAC), combination of cepstral feature with sinusoidal model-based feature (PAMFCC+NHA+HAR+PHF) and combination of temporal, sinusoidal model-based and cepstral features (VTC+PSR+PAMFCC+NHA+HAR+PHF). The comparison of result shows that the combined temporal and sinusoidal model-based (VTC+PSR+NHA+HAR+PHF) feature performs better compared to the individual (VTC+PSR and NHA+HAR+PHF) feature for all three vowels. The result also shows that combined (VTC+PSR+PAMFCC) performs better compared to individual features for /a/ and /u/ vowels and combined (VTC+PSR+SMAC) feature perform better for all three vowels. However, the performance of combined (PAMFCC+NHA+HAR+PHF) is poor compared to the individual PAMFCC feature for all three vowels. The performance of com-

bined (VTC+PSR+PAMFCC+NHA+HAR+PHF) is also poor compared to PAMFCC feature for /a/ and /i/ vowel but it is better for /u/ vowels. These results show that the temporal and sinusoidal model-based features are complementary to each other as the former is temporal and later is a spectral feature. This also shows that the cepstral feature and spectral moment features are also complementary to temporal features for the same reason. However, the performance of combined (VTC+PSR+PAMFCC+NHA+HAR+PHF) feature shows that the three temporal, sinusoidal model-based and cepstral features are not complementary. It can be observed from the results that among individual features PAMFCC and among combined feature its combination with temporal feature, (VTC+PSR+PAMFCC) is performing best for /a/ and /i/ vowels. But it is individual feature SMAC and combined feature VTC+PSR+SMAC which performs best for the case of /u/ vowel. McNemars statistical test also shows the statistically significant ($p < 0.001$) corresponding to these features compared to others.

Table 5.5: Hypernasality detection using combined features for /*a*/ vowel.

Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Acc (%)**	**Sen (%)**	**Spe (%)**
VTC+PSR	71.40±0.12	63.83±4.31	76.48±2.80	79.11±1.32	65.24±5.07	88.67±3.70
NHA+HAR+PHF	81.54±1.97	83.29±4.27	80.39±1.04	82.46±2.30	82.38±5.52	82.45±1.19
PAMFCC	85.23±1.15	84.58±4.59	85.71±3.55	89.56±1.37	89.88±6.19	89.39±4.54
SMAC	80.08±1.16	78.38±3.60	81.24±1.78	84.46±2.07	81.15±5.88	86.78±2.37
MFCC	78.79±1.53	79.61±3.36	78.25±2.98	83.51±1.97	80.58±4.31	85.55±3.83
VTC+PSR+NHA+HAR+PHF	83.77±1.87	85.78±3.51	82.46±1.12	85.37±2.03	86.27±3.36	84.79±1.50
VTC+PSR+PAMFCC	85.70±0.77	85.36±4.49	85.95±3.26	90.68±0.77	90.84±6.89	90.61±3.82
VTC+PSR+SMAC	81.28±0.99	80.16±2.49	82.04±1.78	86.96±1.93	84.35±5.71	88.79±0.82
PAMFCC+NHA+HAR+PHF	86.61±1.13	89.20±5.13	84.89±3.17	88.63±1.41	88.88±6.41	88.50±3.36
VTC+PSR+NHA+HAR+PHF+PAMFCC	86.23±1.26	88.35±4.61	84.83±3.06	88.16±1.38	88.21±5.78	88.16±3.51

Table 5.6: Hypernasality detection using combined features for /*i*/ vowel.

Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Acc (%)**	**Sen (%)**	**Spe (%)**
VTC+PSR	76.57±0.87	77.54±1.33	75.52±2.73	82.74±1.69	83.97±0.67	81.03±2.98
NHA+HAR+PHF	86.38±0.84	84.28±2.32	88.52±1.04	87.89±0.49	84.86±1.13	91.34±1.12
PAMFCC	87.07±2.47	85.16±7.71	88.80±2.58	92.30±1.31	88.93±2.87	95.91±0.66
SMAC	85.83±0.67	81.51±2.64	90.38±2.67	89.91±1.99	84.77±3.26	94.58±3.69
MFCC	84.68±0.56	83.74±1.98	85.29±1.27	89.87±0.86	87.18±2.64	92.76±1.85
VTC+PSR+NHA+HAR+PHF	87.56±1.74	84.22±3.62	90.99±1.12	88.29±2.36	83.84±4.11	93.39±1.11
VTC+PSR+PAMFCC	86.56±1.50	84.97±5.42	88.05±2.31	91.89±2.32	88.04±6.21	96.07±1.52
VTC+PSR+SMAC	87.57±0.92	83.42±2.56	92.01±2.54	91.19±2.92	85.97±4.73	97.02±2.49
PAMFCC+NHA+HAR+PHF	90.99±1.47	87.47±4.04	94.84±1.80	92.03±1.57	88.57±4.24	95.98±2.22
VTC+PSR+NHA+HAR+PHF+PAMFCC	90.54±1.93	86.50±4.62	94.91±1.06	91.51±2.19	86.33±5.29	97.43±1.78

Table 5.7: Hypernasality detection using combined features for /u/ vowel.

Frame level performance				Vowel level performance		
Feature	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Acc (%)**	**Sen (%)**	**Spe (%)**
VTC+PSR	80.35±1.38	84.83±3.35	76.62±0.48	84.44±2.14	88.00±5.97	81.60±1.46
NHA+HAR+PHF	82.65±0.59	85.38±2.15	80.49±1.39	84.25±1.18	86.40±3.45	82.59±2.06
PAMFCC	86.33±2.30	85.91±6.25	86.67±1.12	87.55±2.47	84.43±6.37	90.07±0.79
SMAC	86.07±0.81	90.24±3.84	82.62±2.15	89.19±1.04	92.99±3.02	86.13±3.72
MFCC	85.09±2.16	91.14±5.67	80.08±1.90	87.19±1.19	93.00±2.86	82.53±1.79
VTC+PSR+NHA+HAR+PHF	83.31±0.97	87.02±2.58	80.38±1.66	85.69±1.25	86.55±3.52	85.04±2.06
VTC+PSR+PAMFCC	89.51±1.47	91.30±4.03	88.00±1.05	93.55±2.04	94.92±4.96	92.41±0.49
VTC+PSR+SMAC	89.19±1.19	91.09±4.56	87.57±1.86	93.03±1.15	92.30±4.71	93.60±1.98
PAMFCC+NHA+HAR+PHF	84.79±1.23	90.44±1.99	80.06±1.86	85.70±1.58	90.48±1.44	81.87±2.28
VTC+PSR+NHA+HAR+PHF+PAMFCC	90.99±1.27	95.54±2.73	87.12±1.20	92.61±1.61	95.67±3.62	90.13±0.87

5.7 Summary

In this chapter of the thesis, three spectral features namely HNGDF, PAMFCC and SMAC are used for the hypernasality detection and the result is compared with the MFCC feature. The HNGDF feature is extracted from the HNGD spectrum which is derived from the phase spectrum of the windowed speech. The size of window is taken less than a pitch period which gives the vocal tract characteristics of hypernasal speech within pitch period and the group delay function ensures the high spectral resolution to resolve nasal and oral formants. The PAMFCC feature is extracted from the cepstral smooth spectrum rather than the magnitude spectrum. The pitch-adaptive low time liftering is done to compute the cepstral speech spectrum which eliminates the pitch harmonic effect. The SMAC feature captures the nasal formant and antiformants, oral formants in the hypernasal vowel spectrum along with the spectral envelope information under the notion of pyknogram. It is done by computing the spectral moments of bandpass filtered speech. The HNGDF feature gives poor performance, whereas the PAMFCC and SMAC feature give better performance compared to the MFCC feature for /a/, /i/ and /u/ vowels. The combination of HNGDF with the MFCC feature enhances the performance compared to MFCC feature for /a/ and /i/ vowels. The best performance is obtained using the PAMFCC feature for /a/ and /i/ vowels and using SMAC feature for /u/ vowels. The chapter also compares the individual and inter combination of temporal, sinusoidal model-based and cepstral features for hypernasality detection. The combined (VTC+PSR+PAMFCC) feature is giving the best performance among combined features for /a/ and /i/ vowels and (VTC+PSR+SMAC) feature is giving the best performance for /u/ vowels.

In this thesis, hypernasality detection in the form of normal vs. hypernasal vowels classification is performed in Chapter 3, Chapter 4 and Chapter 5 using temporal, sinusoidal model-based and cepstral features respectively. This detection is the first step in the development of computer-aided diagnosis systems able to assess the degree of hypernasality in speech, and therefore, provide a valuable tool to monitor the effectiveness of a certain therapy in a non-invasive way. However, the doctors and SLPs are more concerned about the severity grading of nasality for the assessment of hypernasality because severity grading gives better information about the velopharyngeal activity which helps in choosing the right medical treatment for caring hypernasality. So in the next chapter, a method for hypernasality severity detection is proposed and validated. A MATLAB based graphical user interface of the proposed method is also implemented in the next chapter.

6

Computation of nasality score for hypernasality severity detection

Publications

- **A. K. Dubey**, A. Tripathi, S. R. M. Prasanna, and S. Dandapat, "Detection of hypernasality based on vowel space area," *J. Acoust. Soc. Am.* 143(5), EL412-EL417(2018).

Contents

Overview

The objective of this chapter is to firstly, propose a method of hypernasality severity detection based on the nasality score between [0 to 1] corresponding to the speaker's speech. The nasality score is computed using the feature extracted from the triad of /a/, /i/ and /u/ vowels present in the speaker's speech rather than the individual vowels. The nasality score is computed by testing the feature extracted from the speaker's speech against a regression model trained using the feature extracted from the speech having different severity of nasality. Later in this chapter, the proposed method is validated by computing the nasality score corresponding to children's speech with normal, mild and moderate-severe hypernasality. Further, the correlation of these nasality scores with the perceptual scores of children's speech is also computed. The validation is done using PAMFCC, combined (VTC+PSR+PAMFCC), vowel space area (VSA), and the baseline MFCC features. Finally, a MATLAB based graphical user interface (GUI) is developed to compute the nasality score for a speaker's speech.

6.1 Introduction

In the earlier chapters, temporal, sinusoidal model-based and cepstral features are explored for the normal vs. hypernasal speech detection in CP speech. The detection accuracy is reported separately for /a/, /i/ and /u/ vowels at the phoneme level. For a particular feature, the accuracy is found different for three vowels. It is comparatively more for high vowels (/i/ and /u/) than the low vowel /a/. Moreover, different features give different accuracy for a particular vowel. The normal vs. hypernasal speech detection is the first step in the development of computer-aided diagnosis systems able to assess the degree of hypernasality in speech, and therefore, provide a valuable tool to monitor the effectiveness of a certain therapy in a non-invasive way. However, as discussed earlier, the doctors and SLPs are more concerned about the severity grading of nasality for the assessment of hypernasality. This is because severity grading gives better information about the velopharyngeal activity which helps in choosing the right medical treatment for caring hypernasality. But as discussed in Chapter 2, very few works have been attempted in the literature for hypernasality severity grading and most of these works are based on the hypernasality severity detection using multi-class classification. The multi-class classification gives classification accuracy and the confusion matrix showing the percentage of correctly classified phonemes corresponding to each class. Again, the multi-class classification is performed for a particular vowel at a time using a feature. So like the normal vs. hypernasal speech classification, the

performance of multi-class classification of particular vowel also vary for different features, and for a particular feature, performance varies for different vowels. Further, from the confusion matrix, it can be inferred that the speech belongs to which class of nasality severity, but multi-class classification does not give a nasality score corresponding to the speech as given by the nasometer device. The nasometer device gives nasality scores between 0 to 100, where scores closed to 0 represent normal speech and scores closed to 100 present severe hypernasal speech. The nasality score is needed because such kind of score is easy to interpret and even an ordinary person having less technical knowledge can infer about the nasality severity using the score. The nasality score may also be more helpful for the SLPs in evaluating the effect of therapy sessions. In literature, a method to obtain such a nasality score is proposed in [92] by modeling the two extremely opposite classes of speech having minimum and maximum nasality. The oral sentences uttered by normal children were considered for the speech with minimum nasality and nasal sentences uttered by the moderate-severe hypernasal children were considered for the speech with maximum nasality. The oral sentences were considered as a test speech and obtained posterior probabilities using the MFCC feature were considered as the hypernasality scores.

Considering the above-mentioned points, in this chapter, firstly a method for hypernasality severity detection is proposed. In this method, a nasality score between [0 to 1] corresponding to the speaker's speech is computed and based on its value hypernasality severity detection is done. If the nasality score is close to 0, the speech is detected as normal speech and if it is close to 1, speech is detected as severe hypernasal speech. The range of scores for mild and moderate hypernasal speech is decided by the SLPs. The nasality score is computed by testing the feature extracted from the speaker's speech against a regression model trained using the feature extracted from the speech having different severity (normal, mild, moderate and severe) of nasality. The feature is extracted from the triad of /a/, /i/ and /u/ vowels rather than the individual vowel present in the speaker's speech. To validate the proposed method, later in this chapter, nasality scores corresponding to the speech of different hypernasal children and the correlation of nasality scores with their perceptual scores are computed. The nasality score and correlation values are computed individually using PAMFCC, combined (VTC+PSR+PAMFCC), vowel space area (VSA), and the baseline MFCC features extracted from the triad of /a/, /i/ and /u/ vowels. The PAMFCC and combined (VTC+PSR+PAMFCC) features are used because these features have yield better performance for normal vs. hypernasal

speech classification for /a/, /i/ and /u/ vowels compared to other features explored in the previous chapters. VSA feature is defined as the two-dimensional area bounded by lines joining the points with coordinates represented by the first F_1 and F_2 of different vowels. The VSA feature is used because hypernasality introduces nasal formant and antiformant pairs in the vowel spectrum which results in the shifting of formants. This shifting may affect the size of the VSA as the severity of nasality increases. The VSA has been used to study the characteristics of various pathology like psychological distress [157], hearing impairment [158] and also for analyzing the regional dialect variation and sound change [159]. In the last part of this chapter, using the concept of the proposed method a MATLAB based graphical user interface (GUI) is developed to compute the nasality score for a speaker's speech.

The rest of this chapter is organized as follows: Section 6.2 discusses the proposed method for hypernasality severity detection. Section 6.3 discusses the validation of the proposed system by computing the nasality scores and their correlation with perceptual scores for different severity of hypernasal speech. Section 6.4 discusses the implementation of proposed method for clinical application. Finally the Section 6.5, gives the summary and conclusion of the presented work.

6.2 Proposed method for hypernasality severity detection

In this proposed method for hypernasality severity detection, the nasality score corresponding to the speaker's speech is computed. The nasality score is computed for each triad of /a/, /i/ and /u/ vowels present in the speaker's speech and mean of all scores are taken to find the nasality score for speaker's speech. Fig.6.1 shows the block diagram of the process followed to compute the nasality score for the triad of /a/, /i/ and /u/ vowels. The whole process of nasality score computation has training and testing phases. The speech data used for the method are the /CVCV/ words consisting of /a/, /i/ and /u/ vowels. In the preprocessing stage of both the phases, vowels /a/, /i/ and /u/ are detected from the /CVCV/ speech utterances using the manual annotation or some signal processing technique. In the next step, the feature is extracted from each triad of /a/, /i/ and /u/ vowels, both in training and testing phases. It is done by first computing feature at the phoneme level from the three vowels and then concatenating them in a vector. The computation of feature at the phoneme level is done by first computing the feature at the frame level and taking the mean of feature corresponding to all frames present in the phoneme. In the training phase, a regression model is trained with the feature extracted from vowels corresponding to normal, mild, moderate and severe hypernasal speech. In the

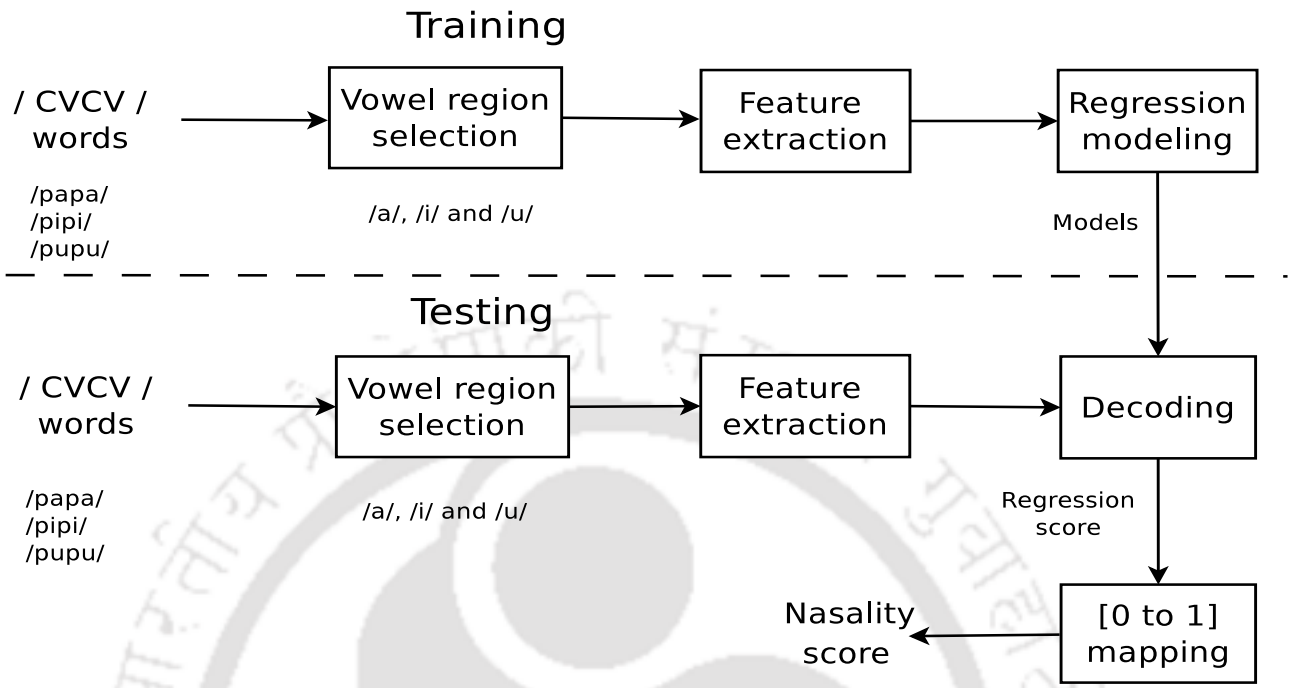

Figure 6.1: Block diagram of the proposed system for the hypernasality assessment based on the severity grading of speech.

testing phase, the feature extracted from vowels present in a test speaker's speech is tested against the trained regression models. The regression score obtained for each triad of vowels is mapped between [0 to 1] and it is considered as the nasality score. The mean of nasality scores for all triads of vowels present in the test speaker's speech is taken to denote the nasality for that test speaker's speech. The obtained nasality score may supplement the decision of the doctors and SLPs for hypernasality severity grading of the speaker's speech. This can be done by setting the threshold levels of nasality score for normal, mild, moderate and severe hypernasal speech. In a very general way, an equal interval of nasality score, 0-0.25 for normal, 0.26-0.50 for mild, 0.51-0.75 for moderate and 0.76-1 for severe hypernasal speech, can be considered. However, this range can be decided by the SLPs based on the age, gender, and language of the test speaker.

6.3 Validation of proposed system

To validate the effectiveness of nasality score proposed for hypernasality severity detection, in this section the nasality scores corresponding to the 15 children with normal speech, 15 children with mild hypernasal speech, and 15 children with moderate-severe (10 moderate and 5 severe) hypernasal speech and the correlation of these scores with the children's perceptual score are computed. The children

are same as presented in Table 3.2 of Chapter 3 (Section 3.2). As the numbers of children with normal and mild hypernasal speech are 15, but the numbers of children with moderate and severe hypernasal speech are 10 and 5, respectively, so the number of children with moderate and severe hypernasal speech is taken in a single group with the moderate-severe hypernasal speech. To compute nasality score, 99 normal, 91 mild, 101 moderate-severe (71 moderate and 30 severe) samples of each /papa/, /pipi/ and /pupu/ words were taken from these children's recorded speech. As one /CVCV/ word contains two vowels, so the nasality scores corresponding to these children's speech is computed using the 198 normal, 182 mild, 202 moderate-severe triads of /a/, /i/ and /u/ vowels. The glottal activity detection algorithm proposed in [160] is used to detect the vowels from the /CVCV/ words. The algorithm classifies the voiced and unvoiced region in speech using the combination of the strength of excitation, normalized autocorrelation peak strength, and higher-order statistics features. Table 3.2 has already shown the severity rating agreement between each pair of SLP these children's speech.

6.3.1 Feature extraction

The nasality score corresponding to children's speech is computed individually using PAMFCC, combined (VTC+PSR+PAMFCC), VSA and baseline MFCC features. The features are extracted for each triad of /a/, /i/ and /u/ vowels using the procedure explained below.

- **PAMFCC feature:** To compute the PAMFCC feature for a triad of /a/, /i/ and /u/ vowel, the feature is first computed separately for all three vowels at the phoneme level and then all are concatenated in a vector. The computation of feature at the phoneme level is done by computing the feature at the frame level and taking the means of feature of all frames present in the phoneme. The procedure to compute the PAMFCC feature from a frame is the same as explained in Chapter 5. As the PAMFCC feature is a 13-dimensional for a frame of speech, so the dimension of PAMFCC feature for a triad will be 39-dimensional.

- **VTC+PSR+PAMFCC feature:** The VTC+PSR+PAMFCC feature for a triad of /a/, /i/ and /u/ vowel is computed similar way as PAMFCC feature. The procedure to compute the VTC+PSR feature for a frame of speech is computed as explained in Chapter 3. As the VTC+PSR+PAMFCC feature is a 15-dimensional (2-VTC+PSR+ 13-PAMFCC) for a frame of speech, so the dimension of VTC+PSR+PAMFCC feature for a triad will be 45-dimensional.

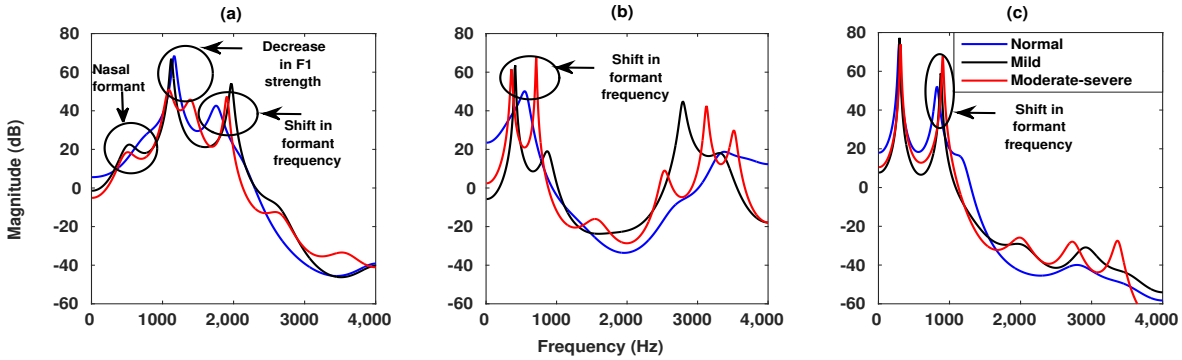

Figure 6.2: Linear Prediction spectrum of the normal, mild and moderate-severe hypernasal vowels. (a) for /a/ vowel (b) for /i/ vowel (c) for /u/ vowel. Figure shows the addition of extra formants, reduction in formant strength and shift in formant frequency for hypernasal vowels.

- **VSA feature:** VSA is defined as the area of the shape formed by lines joining the points with F_1 and F_2 frequencies of different vowels as coordinates. As in this work, three vowels /a/, /i/ and /u/ are considered, so here VSA is the area of a triangle. For computation of VSA feature for a triad of /a/, /i/ and /u/ vowel, F_1, F_2 frequencies of each frame of these vowels are computed and mean of these frequencies are taken to find the F_1, F_2 frequencies at the phoneme level. The frame size of 20 ms, frameshift of 10 ms and LP analysis is used to compute formant frequencies (F_1, F_2) of three vowels which form the coordinates of three vertices of the vowel triangle. Euclidean distance between each pair of vertices gives the length of three sides of the triangle. Area of the triangle is calculated using the Herons formula

$$Area = \sqrt{s(s-a)(s-b)(s-c)} \tag{6.1}$$

where $s = \frac{a+b+c}{2}$ and a, b, c are the lengths of three sides of the triangle. The presence of nasal formant-antiformant pairs in the spectrum of hypernasal vowels reduce the formants strength and shift the locations of F_1 and F_2 frequencies. Fig. 6.2 (a)-(c) shows the linear prediction spectrum of normal, mild and moderate-severe hypernasal /a/, /i/ and /u/ vowels respectively. From Fig. 6.2, the addition of extra nasal formant around F_1, reduction in F_1 strength and shift in formant frequency in hypernasal vowels compared to normal can be observed. This shifting of formants affects the VSA of hypernasal speech as the severity of nasality increases.

To show the nature of VSA feature for normal, mild and moderate-severe hypernasal speech, the F_1 and F_2 frequencies of all /a/, /i/ and /u/ vowels corresponding to normal, mild and moderate-

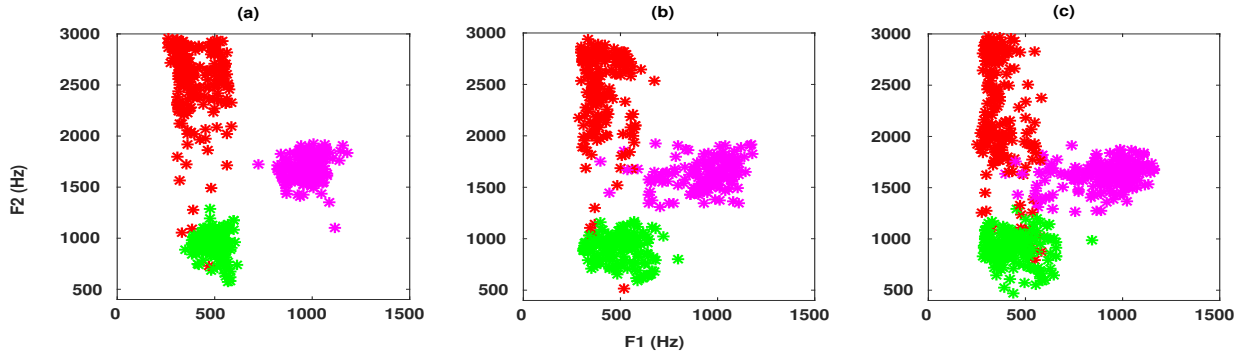

Figure 6.3: Vowel space plots of (a) normal, (b) mild and (c) moderate-severe hypernasal speech.

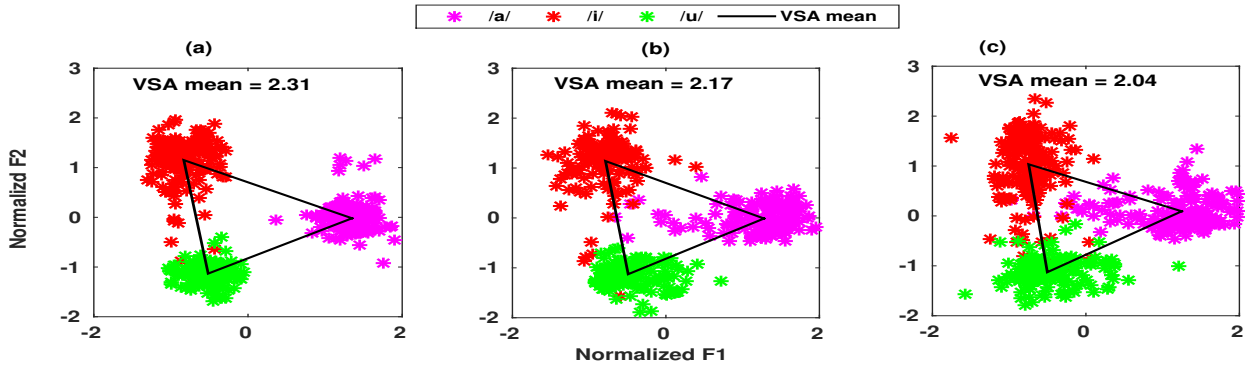

Figure 6.4: Normalized vowel space plots of (a) normal, (b) mild and (c) moderate-severe hypernasal speech.

severe hypernasal speech are computed. Fig. 6.3 shows the 2-D space plot of these frequencies by considering (F_1, F_2) frequencies as the coordinate. Fig. 6.3 (a)-(c) are plotted for normal, mild and moderate-severe hypernasal speech respectively. From Fig. 6.3 it can be observed that there is a high variance in the F_2 values of /i/ vowel which can affect the VSA values. This variation may be because of speaker variability. Hence, each frequency extracted from all three vowels is z-score normalized at the speaker level. The normalized (F_1, F_2) frequencies are plotted in 2-D space as shown in Fig. 6.4 (a)-(c) for normal, mild and moderate-severe hypernasal speech respectively. Fig. 6.4 it can be observed that the variance in the F_2 values of /i/ vowel has reduced. From Fig. 6.4, it can also be observed that the regions of formant frequencies coordinates are coming closer to each other as the severity of hypernasality increases from normal through mild to moderate to severe. Fig. 6.4 also shows the vowel triangles formed by the mean values of all formant frequency coordinates of /a/, /i/ and /u/ vowels and it is called as VSA mean. The calculated VSA mean is 2.31 for normal speech, 2.17 for mild hypernasal

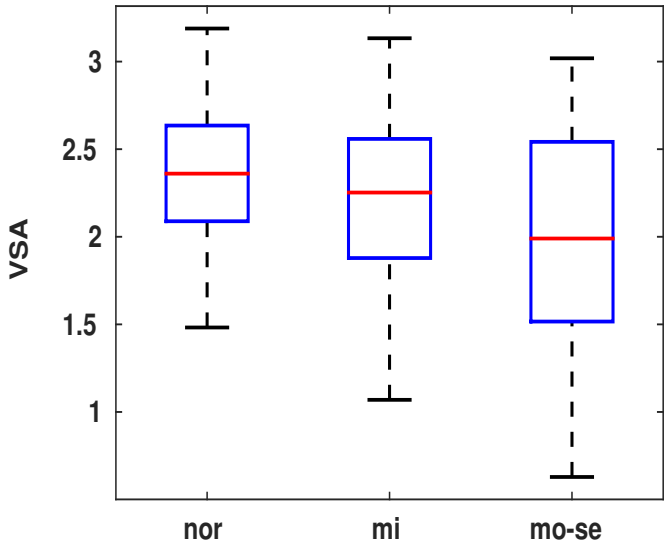

Figure 6.5: Box plot showing the nature of VSA feature for normal, mild and moderate-severe hypernasal speech. nor=normal, mi=mild and mo-se=moderate-severe.

speech and 2.04 for moderate-severe hypernasal speech. This indicates that the VSA mean decreases as the severity of hypernasality increases. To compare the difference in VSA values computed for all 198 normal, 182 mild, 202 moderate-severe triads of /a/, /i/ and /u/ vowels, the boxplot of VSA values is plotted in Fig. 6.5. It can be observed that the median of VSA decreases as the severity of hypernasality increases.

- **MFCC feature:** The computation of baseline MFCC feature for a triad of vowels is similar to the PAMFCC feature and its dimension is also 39.

6.3.2 Computation of nasality score and correlation value

The computed features from 198 normal, 182 mild, 202 moderate-severe triads of /a/, /i/ and /u/ vowels are used to compute the nasality score corresponding to 15 children with normal speech, 15 children with mild hypernasal speech and 15 children with moderate-severe hypernasal speech. To do the assessment, the nasality score for each child is computed. The assessment is done separately using PAMFCC, VTC+PSR+PAMFCC, VSA and MFCC features. To compute nasality scores, the leave-one-speaker-out procedure of training and testing is followed. Since there is a total of 45 speakers, so 45 linear regression models are built for all the folds. The mean of scores corresponding to all triads of /a/, /i/ and /u/ vowels in a speaker's speech is the nasality score for that speaker. Fig. 6.6 (a)-

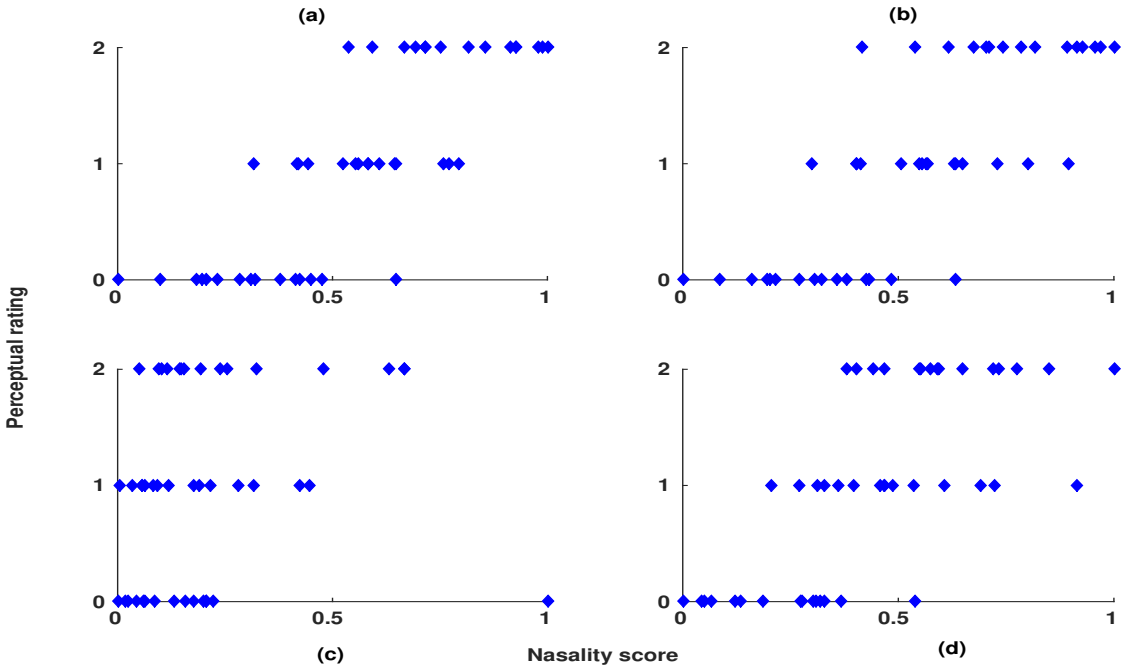

Figure 6.6: Scatter plots of nasality scores obtained separately using (a) PAMFCC feature (b) VTC+PSR+PAMFCC feature (c) VSA feature and (d) baseline MFCC feature. The scores are computed for the children having different severity of nasality in their speech. Here 0, 1, 2 perceptual rating represents normal, mild and moderate-severe hypernasality respectively.

(d) shows the nasality scores for all 15 children with normal speech, 15 children with mild hypernasal speech and 15 children with moderate-severe hypernasal speech for PAMFCC, VTC+PSR+PAMFCC, VSA, and MFCC features respectively. The perceptual ratings 0, 1, 2 respectively represent normal, mild and moderate-severe hypernasality. So, at the perceptual score 0, the nasality scores of 15 children with normal hypernasal speech are plotted. At the perceptual score 1, the nasality scores of 15 children with mild hypernasal speech are plotted. At the perceptual score 2, the nasality scores of 15 children with moderate-severe hypernasal speech are plotted. The nasality score is between 0 to 1, where, 0 represents normal speech and 1 represents severe hypernasality. From Fig. 6.6 it can be observed that the nasality score is closed to 0 for normal speech, around 0.5 for mild and closed to 1 for moderate-severe hypernasal speech. It can also be observed that compared to the MFCC feature, PAMFCC and VTC+PSR+PAMFCC features are giving better discrimination among the nasality scores corresponding to normal, mild and moderate-severe hypernasal speech, but VSA feature is giving poor discrimination compared to the MFCC feature.

Table 6.1: Correlation of nasality scores with the perceptual scores and p-value for different features

Feature	Correlation coefficient	p-value
PAMFCC	0.82	<0.001
VTC+PSR+PAMFCC	0.78	<0.001
VSA	0.30	0.05
MFCC	0.73	<0.001
VSA+MFCC	0.74	<0.001
VTC+PSR+PAMFCC+VSA	0.78	<0.001
PAMFCC+VSA	0.82	<0.001
Nasometer reading	0.78	<0.001

6.3.3 Computation of correlation value

To prove the observation in Fig. 6.6, correlation of nasality score obtained corresponding to these features with the perceptual rating is computed. Table 6.1 shows the correlation coefficient and p-value corresponding to different feature. The correlation value is high for PAMFCC and VTC+PSR+MFCC features compared to the MFCC feature, but it is low for the VSA feature. The value is highest for the PAMFCC feature. The correlation is also computed for a combination of VSA feature with the MFCC, VTC+PSR+PAMFCC and PAMFCC feature. The combined VSA+MFCC feature shows the improvement in the correlation value over the value corresponding to the MFCC feature. This shows that the VSA feature alone may not be effective for the discrimination of normal, mild and moderate-severe hypernasality, but its combination with the MFCC can be used for the hypernasality severity detection. However, the improvement in correlation value is not obtained for the combined VSA+PAMFCC and VSA+VTC+PSR+PAMFCC features. This experiment using different features shows that the nasality scores are capable of hypernasality severity detection. The best correlation is obtained for The PAMFCC feature. Table 6.1 also shows the correlation of perceptual rating with the nasometer reading, which is used in the clinical environment for hypernasality severity detection. It can be observed from Table 6.1 that the correlation of PAMFCC feature is better compared to nasometer reading. So, the nasality scores obtained corresponding to PAMFCC feature can be used by the SLPs to supplement their perceptual decision for hypernasality severity detection.

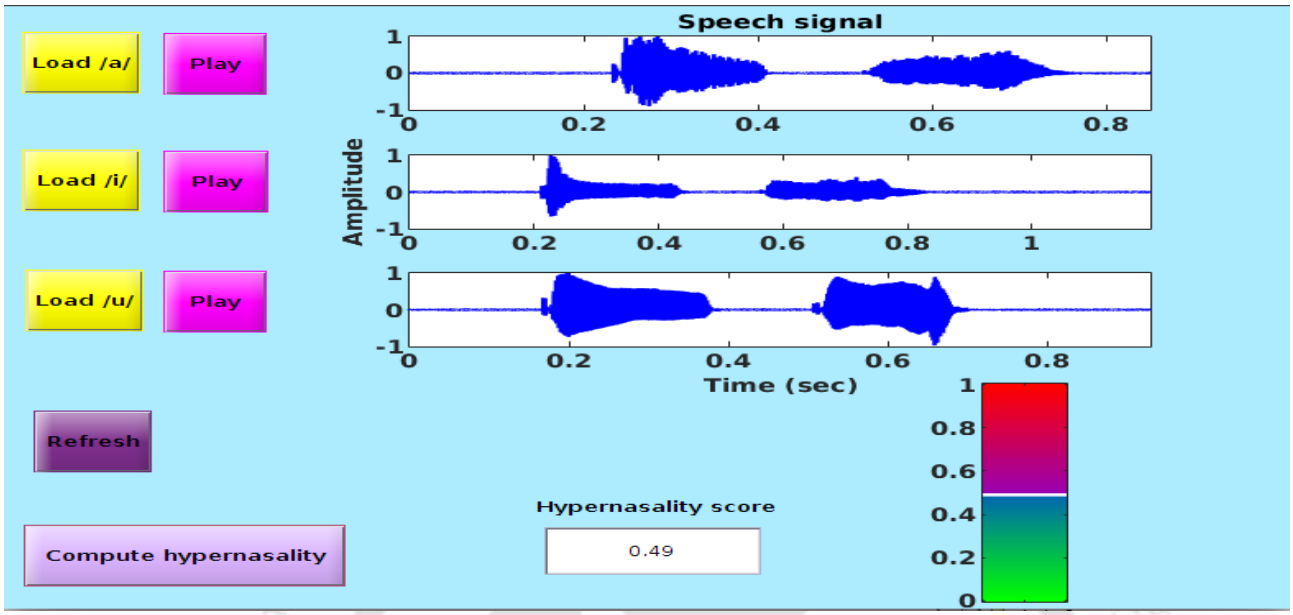

Figure 6.7: Screenshot of GUI showing the hypernasality score using PAMFCC feature

6.4 Implementation of proposed method for clinical application

In this section, the description of MATLAB based GUI implementation of the proposed method of hypernasality severity detection is discussed. The GUI could be used for the clinical application to display the nasality score corresponding to CP children's speech. Fig. 6.7 shows the screenshot of GUI which has three load buttons to load the /papa/, /pipi/ and /pupu/ words uttered by a child with CP. The GUI also has three play buttons to play and listen to the uploaded .wav files. The uploaded .wav files are displayed on the three display panels. The compute hypernasality button is used to compute the nasality score which is displayed on a display panel. The score is also shown using a colour bar where green colour represents normal hypernasality and red represents moderate-severe hypernasality. In the back end of the GUI, the proposed method of nasality score computation has been implemented. So in the back end of GUI, first of all /a/, /i/ and /u/ vowels are detected form the uploaded /papa/, /pipi/ and /pupu/ words using the glottal activity detection algorithm. This will give the two triads of /a/, /i/ and /u/ vowels for which two nasality scores are computed and mean of these scores are displayed as the nasality score corresponding to speaker's speech. The nasality score is computed after testing the feature against the trained linear regression model. The back end of this GUI has the regression model trained using the feature extracted from the speech of normal, mild and moderate-severe hypernasal speech. Both for training and testing, the feature is

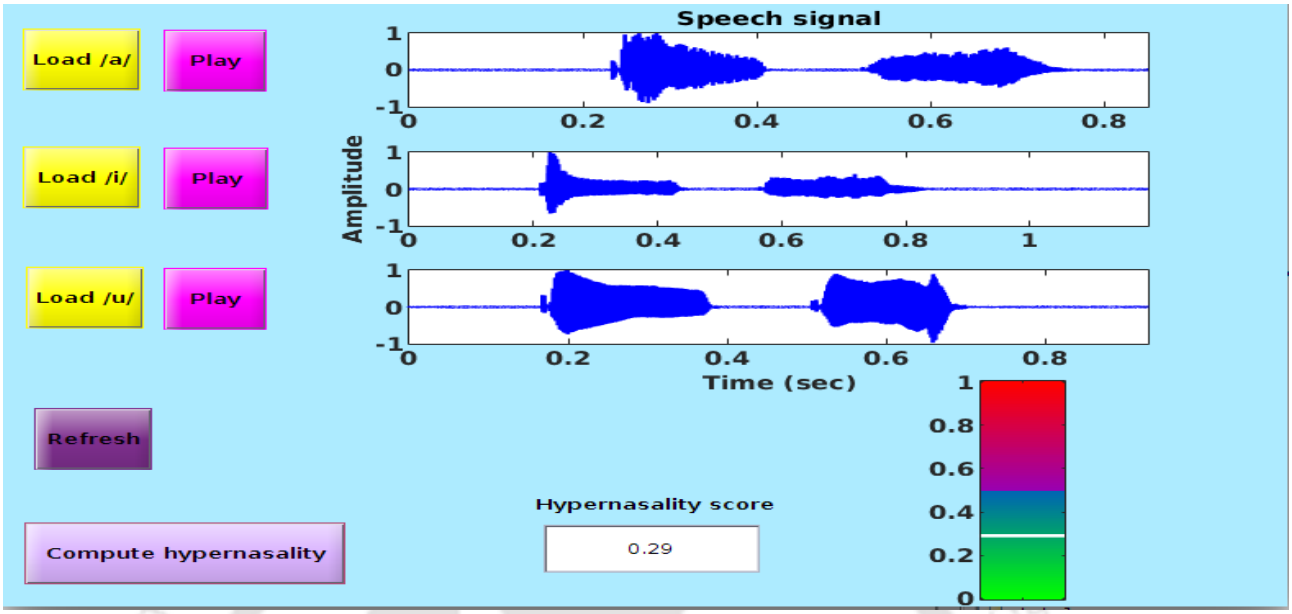

Figure 6.8: Screenshot of GUI showing the hypernasality score using MFCC feature

extracted from the triads of /a/, /i/ and /u/ vowels. The PAMFCC feature is used to develop the GUI as it is giving the best correlation with the perceptual scores as discussed in the previous section. To show the performance of the GUI, the nasality score corresponding to the speech of a child with mild hypernasality is computed using the GUI. Fig. 6.7 shows the screenshot of GUI showing the nasality score using the PAMFCC feature. The nasality score of 0.49 is shown on the display panel. For the comparison, the nasality score corresponding to same child's speech is computed using the GUI developed using MFCC feature. Fig. 6.8 shows the screenshot of GUI showing the nasality score using the MFCC feature. The nasality score of 0.29 is shown on the display panel. The nasality score of 0.29 may be confusing for the SLPs to declare the speech as mild but with the score of 0.49, SLPs can come to the conclusion that the speech is mild. This PAMFCC based GUI could be validated for the clinical application and SLPs could use this GUI to supplement their perceptual rating about the hypernasal speech produced by children with CP.

6.5 Summary

In this chapter, a method for hypernasality severity detection is proposed. In this method nasality score between [0 to 1] corresponding to the speaker's speech is computed. The computed nasality score could be used by the SLPs for the severity detection of hypernasality present in the speaker's

speech based on the threshold values for normal, mild, moderate and severe hypernasal speech. In this method, a linear regression model is trained using the feature computed from the normal, mild, moderate and severe hypernasal speech and in the testing phase feature computed from the test speaker's speech is tested against the trained regression model. The regression scores are mapped between [0 to 1] to obtain the nasality score. The feature is computed from the triad of /a/, /i/ and /u/ vowels rather than the individual vowels. The proposed method was validated in this chapter by computing the nasality score corresponding to the 15 children with normal, 15 with mild and 15 with moderate-severe hypernasal speech and correlation of these scores with the perceptual scores. The validation is done using PAMFCC, VTC+PSR+PAMFCC, VSA and MFCC features. The experiment shows that the best correlation of 0.82 is obtained for the PAMFCC feature. This experiment shows that the proposed method could be used to supplement the perceptual assessment of hypernasality in CP children in the clinical environment. Finally, the proposed method is implemented in the form of a MATLAB based GUI for the clinical application.

In this thesis attempts are made for the hypernasality detection in CP children's speech. The temporal, sinusoidal model-based and cepstral features are explored for hypernasality detection in chapter 3, 4 and 5 respectively. In this chapter, a method for hypernasality severity detection is proposed and validated. Next chapter provides the summary, contributions of the present thesis and the directions for future work.

7

Conclusions

Contents

7.1 Summary of the work

This thesis focused on the hypernasality detection and assessment of hypernasality in CP speech using novel features. The hypernasality assessment can help doctors and SLPs in choosing the right medical treatment for children having hypernasal speech. The proposed method is based on the hypernasality severity detection of /a/, /i/ and /u/ vowels present in CP speech. The vowels are used because vowels get nasalized in hypernasal speech produced by CP children. The spectral analysis of hypernasal vowels shows the presence of additional nasal formant and antiformat pairs in the spectrum which centralizes the energy in the lower frequencies, affects the strength of harmonics along with the spectral envelope. So, there is a deviation in the spectral characteristics of hypernasal vowels as the severity of hypernasality increase from normal to mild to moderate-severe. In this thesis, firstly, normal vs. hypernasal vowel detection was performed to explore features that are capable of capturing the spectral deviation present in hypernasal speech. The hypernasality detection was done using the temporal features, sinusoidal model-based features, cepstral features, and their inter combinations. The temporal features capture the effect of nasal formant and antiformant pairs on the temporal characteristics of speech waveform and LP residual signal of hypernasal vowels. The sinusoidal model-based features capture the effect of nasal formant and antiformant pairs on the strength of harmonics and frequency location of prominent harmonics. The cepstral features capture the effect of nasal formant and antiformant pairs on the spectral envelope. Later, hypernasality severity detection was performed based on the nasality score between [0 to 1] corresponding to the speaker's speech. The nasality scores are computed using the best performing feature for normal vs. hypernasal vowel detection and VSA feature. In the last, a MATLAB based GUI to compute the nasality score is developed which could be used for the assessment of hypernasality in CP children.

The important results of the hypernasality detection done in the thesis are as follows:

- **Hypernasality detection using temporal features:** In the first work, hypernasality detection is attempted using the VTC and PSR features. The VTC feature is extracted from the speech signal to capture the low-frequency prominence in the signal. The presence of nasal formant in hypernasal vowel spectrum enhances the low-frequency harmonics. The value of VTC feature is high for hypernasal speech compared to normal speech. The PSR feature is extracted from the LP residual of the speech signal to capture the residual signal characteristics around epoch locations. The presence of nasal format and antiformant pairs in hypernasal speech adds

undesirable signal component in the residual signal around the epoch locations. The value of PSR feature is low for hypernasal speech compared to normal speech. Both the features are the temporal feature and do not involve low-pass filtering of the speech with a predefined cutoff frequency as requires by most of the baseline features. The box plots for VTC and PSR features show their distinctive nature for normal and hypernasal vowels. The result of ANOVA test shows the statistical significance of both the features for normal and hypernasal vowels. Hence two-dimensional (VTC+PSR) feature is proposed for normal vs. hypernasal vowel classification using the SVM classifier. The classification performance for the proposed feature is better compared to the baseline features for each vowel, but it is poor compared to the performance of the MFCC feature.

- **Hypernasality detection using sinusoidal model-based features:** In the second work, detection of hypernasality is done using the sinusoidal model-based NHA, HAR, and PHF features. The features are based on the strength of harmonics in the spectrum which is measured using the sinusoidal model of speech. The NHA feature is the magnitude of harmonics with respect to their maximum magnitude. The HAR feature is the relative magnitude of a harmonic with respect to its previous harmonic magnitude. The PHF feature is the frequency location of prominent harmonics in the spectrum. The analysis shows that the nature of these features is different for hypernasal vowels compared to normal vowels. It happens because the presence of nasal formant and antiformant pairs in the hypernasal vowels spectrum directly modifies the harmonic strength. To measure the discriminative capability of each dimension of the features, a statistical SD measure between feature dimensions and class labels are used, and the feature dimensions are arranged in the decreasing order of SD measure. The high SD value of feature dimension shows the high discriminative capability for normal and hypernasal vowels. This point is demonstrated by the normalized histogram plots and by comparing the mean and std values for normal and hypernasal vowels. The classification performance of combined (NHA+HAR+PHF) feature is better compared to the (VTC+PSR) feature for /a/ and /i/ vowels and nearly equal for /u/ vowel. However the performance is poor compared to MFCC feature.

- **Hypernasality detection using cepstral features:** In the third work, three cepstral features namely, HNGDF, PAMFCC, and SMAC features are used for hypernasality detection. The performances of the three features are compared with the MFCC feature. The HNGDF feature

is extracted from a high spectro-temporal HNGD spectrum which is derived from the phase spectrum of the ZTW speech. The size of window to compute HNGD spectrum is taken around a pitch period to capture the vocal tract characteristics within a pitch period. The high spectral resolution of HNGD spectrum is due to the group delay function and it resolves the closely spaced nasal and oral formant in the hypernasal vowels. The PAMFCC feature is extracted from the cepstral smooth spectrum rather than the magnitude spectrum. The pitch-adaptive low time liftering is done to compute the cepstral smooth spectrum which is free from the pitch harmonic effect. The PAMFCC feature captures nasality evidence present in the lower frequencies in a better way and gives better performance compared to the MFCC feature. The SMAC feature is the concatenation of spectral moments of signals obtained after band pass filtering of speech signal with the lower order cepstral coefficients. The feature captures the nasal formants and antiformants, oral formants and spectral envelope information in the spectrum under the notion of pyknogram. Compared to MFCC feature, the HNGDF feature gives poor performance for hypernasality detection, whereas the PAMFCC and SMAC feature give better performance for /a/, /i/ and /u/ vowels. The combination of HNGDF with the MFCC feature enhances the performance compared to the MFCC feature for /a/ and /i/ vowels. In the last, hypernasality detection is also performed using the inter combination (at the score level) of temporal feature, sinusoidal model-based feature and cepstral feature. The result shows that the combination of temporal feature with the cepstral feature is giving the best performance. The combined feature (VTC+PSR+PAMFCC) performs best for /a/ and /i/ vowels, whereas combined feature (VTC+PSR+SMAC) performs best for /u/ vowel.

- **Hypernasality severity detection using nasality score:** In the last work, hypernasality severity detection is attempted because it helps the doctors and SLPs in performing the hypernasality assessment. The severity detection is done one three-point scale: normal, mild and moderate-severe hypernasal speech using a nasality score between [0 to 1]. The nasality score is computed corresponding to each triad of /a/, /i/ and /u/ vowels present in the speech and the mean of all scores is taken to give the score at the speaker level. The nasality score is computed using the PAMFCC feature, (VTC+PSR+PAMFCC) feature, VSA feature and baseline MFCC feature. The correlation of nasality score and the perceptual score is computed to measure the usefulness of nasality score for hypernasality assessment. Finally, the proposed hypernasality

severity detection method is implemented in the form of a MATLAB based GUI for the clinical application.

7.2 Discussion

In this thesis, novel features for detection of hypernasality in cleft palate speech are explored, and using these features a system for rating the scale of hypernasality in speaker's speech is proposed. As to my best knowledge, there is no publicly available database for hypernasal speech in CP speech, so one such database at the word-level is collected in collaboration with the SLPs of All India Institute of Speech and Hearing (AIISH), Mysuru, India. In literature, temporal, formant-based, and cepstral coefficient features have been explored for hypernasality detection. However, most of the temporal features have the limitation that the features require the low pass filtering of the speech with a predefined cutoff frequency above the F_1. This is done to capture the centralized energy in the low-frequency region of the spectrum due to the presence of nasal formant and antiformant. But the value of cutoff frequency is different for the different vowels, and for a particular vowel also, the variation in cutoff frequency affects the detection accuracy. Also, the temporal features are only extracted from the speech signal. However, the presence of nasal formant and antiformant in the spectrum may also affect the temporal characteristics of the LP residual signal. The residual signal is never explored for hypernasality detection. So in this thesis temporal features, which do not require the low pass filtering of the speech with a predefined cutoff frequency, are explored from the speech as well as from the residual signal. The formant analysis based features have the limitations that a high-resolution spectrum is needed to resolve the nasal and oral format and there may be spurious peaks in the spectrum which do not represent the formant. Also, finding the exact location and strength of formants in high pitch children's speech is not considered an easy job. So in this thesis, instead of measuring the nasal and oral formant strength in lower frequencies, the strength of harmonics in lower frequencies is explored using the sinusoidal model of speech, and features are proposed for the hypernasality detection. The issues with the explored cepstral features such as MFCC or LPCC are that these features are explored from the DFT magnitude spectrum and phase spectrum is ignored. Further, the DFT magnitude spectrum may do not have enough resolution to resolve the closely spaced nasal and oral tract formant in hypernasal speech. The magnitude spectrum also has the pitch harmonics effect which causes the high variance for the higher coefficients of the MFCC feature. The

high variance affects hypernasality detection accuracy. To overcome these issues new cepstral features extracted from the phase spectrum, cepstral feature free from pitch harmonics effect, and cepstral coefficient combined with spectral moment based feature are explored in this thesis for hypernasality detection. The limitation with the hypernasality severity detection based on multi-class classification is that classification does not give a continuous nasality score just like the nasometer device. So in this thesis, a system for obtaining hypernasality severity score corresponding to children's speech is attempted. The hypernasality severity detection is attempted using the nasality score between [0 to 1] for the severity grading of normal, mild, and moderate-severe hypernasal speech. The nasality score is computed using features extracted from each triad of /a/, /i/ and /u/ vowels present in the speech. In the last, a MATLAB based GUI of the proposed hypernasality severity detection method is implemented that can be tested for the clinical application. The database used for the experiments reported in this thesis is recorded in a sound-treated room using Bruel & Kjaer sound-level meter microphone (type 2250-s hand-held analyzer). This means that the effect of different recording locations, variations in the microphone and presence of background noise etc. is not studied in this thesis. It is expected the these practical issues will have their effect in the performance of hypernasality detection if the proposed system is tested for the speech recorded in the background noise environment. For that some signal processing method to reduce the effect of noise in the recorded speech, change in recording environment and microphone can be used in the pre-processing step of the proposed system.

7.3 Directions for future work

Based on the outcome of this thesis work, this section provides a few potential directions for future research.

- The segmentation of vowel region in the stimuli words are done by manually marking using wavsurfer tool. The detection of vowels can be attempted using signal processing method for automatic hypernasality detection.

- The doctors and SLPs are more interested in severity grading of hypernasality on 4-point scale, which indicates the velopharyngeal gap size. In this thesis severity grading is not attempted individually for /a/, /i/ and /u/ vowels that can be attempted by training the classifier using the data of signal vowel at a time.

- Hypernasality detection and its severity grading can be better perceived by listening the continuous speech. So in future, database containing recording of sentences like /Buy baby a bib/, having mostly vowels and voiced consonants can be considered for hypernasality detection and severity grading.
- The hypernasality detection in this thesis is attempted using the SVM classifier. The detection can be done using deep neural network (DNN) or convolutional neural network.
- The abnormalities in pitch, loudness and voice quality occurring in conjunction with hypernasality in CP children. Hence in future, their effect on perception of hypernasality can be explored.
- The end-user evaluation of GUI can be tested at the AIISH.
- The speech samples used for the work is recordered in the sound-treated room. However, in practical condition speech samples recorded in noisy condition will have speech added with noise. So in future the performance of different proposed features and system can be tested for the noisy speech.

Bibliography

[1] G. Henningsson, D. P. Kuehn, D. Sell, T. Sweeney, J. E. Trost-Cardamone, and T. L. Whitehill, "Universal parameters for reporting speech outcomes in individuals with cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 45, no. 1, pp. 1–17, 2008.

[2] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70, no. 10, pp. 1741–1747, 2006.

[3] A. W. Kummer and L. Lee, "Evaluation and treatment of resonance disorders," *Language, Speech, and Hearing Services in Schools*, vol. 27, no. 3, pp. 271–281, 1996.

[4] A. W. Kummer, *Cleft palate & craniofacial anomalies: Effects on speech and resonance*. Nelson Education, 2013.

[5] J. Andrews and D. Rutherford, "Contribution of nasally emitted sound to the perception of hypernasality of vowels." *The Cleft Palate Journal*, vol. 9, pp. 147–156, 1972.

[6] N. M. Joy and S. Umesh, "Improving acoustic models in torgo dysarthric speech database," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, pp. 637–645, 2018.

[7] K. Stevens, R. Nickerson, A. Boothroyd, and A. Rollins, "Assessment of nasalization in the speech of deaf children," *Journal of Speech, Language, and Hearing Research*, vol. 19, no. 2, pp. 393–416, 1976.

[8] B. L. Eppley, J. A. van Aalst, A. Robey, R. J. Havlik, and A. M. Sadove, "The spectrum of orofacial clefting," *Plastic and Reconstructive Surgery*, vol. 115, no. 7, pp. 101e–114e, 2005.

[9] R. I. Zraick and J. M. Liss, "A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 4, pp. 979–988, 2000.

[10] K. Bettens, F. L. Wuyts, and K. M. Van Lierde, "Instrumental assessment of velopharyngeal function and resonance: A review," *Journal of Communication Disorders*, vol. 52, pp. 170–183, 2014.

[11] A. Stellzig-Eisenhauer, "The Influence of Cephalometric Parameters on Resonance of Speech in Cleft Lip and Palate Patients An Interdisciplinary Study," *Journal of Orofacial Orthopedics/Fortschritte der Kieferorthopädie*, vol. 62, no. 3, pp. 202–223, 2001.

[12] C. Havstam, A. Lohmander, C. Persson, H. Dotevall, A. Lith, and J. Lilja, "Evaluation of VPI-assessment with videofluoroscopy and nasoendoscopy," *British Journal of Plastic Surgery*, vol. 58, no. 7, pp. 922–931, 2005.

[13] D. J. Lam, J. R. Starr, J. A. Perkins, C. W. Lewis, L. E. Eblen, J. Dunlap, and K. C. Sie, "A comparison of nasendoscopy and multiview videofluoroscopy in assessing velopharyngeal insufficiency," *Otolaryngology-Head and Neck Surgery*, vol. 134, no. 3, pp. 394–402, 2006.

[14] P. D. Witt, J. L. Marsh, E. G. McFarland, and J. E. Riski, "The evolution of velopharyngeal imaging." *Annals of Plastic Surgery*, vol. 45, no. 6, pp. 665–673, 2000.

[15] D. W. Warren, "Nasal emission of air and velopharyngeal function," *The Cleft Palate Journal*, vol. 4, no. 2, pp. 148–156, 1967.

[16] R. Foy, "Contribution rhinométrique à létude de la respiration nasale," *Ann Mal Oreille Larynx Nez Pharynx*, vol. 36, pp. 130–149, 1910.

[17] P. Devani, R. Watts, and A. F. Markus, "Speech outcome in children with cleft palate: aerophonoscope assessment of nasal emission," *Journal of Cranio-Maxillofacial Surgery*, vol. 27, no. 3, pp. 180–186, 1999.

[18] H. Dotevall, H. Ejnell, and B. Bake, "Nasal airflow patterns during the velopharyngeal closing phase in speech in children with and without cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 38, no. 4, pp. 358–373, 2001.

[19] J. Quigley, F. Shiere, R. C. Webster, and C. M. Cobb, "Measuring Palatopharyngeal Comptence with the Nasal Anemometer," *The Cleft Palate Journal*, vol. 1, no. 3, pp. 304–313, 1964.

[20] D. W. Warren and A. B. DuBois, "A pressure-flow technique for measuring velopharyngeal orifice area during continuous speech," *The Cleft Palate Journal*, vol. 1, no. 1, pp. 52–71, 1964.

[21] Y. Horii, "An accelerometric measure as a physical correlate of perceived hypernasality in speech," *Journal of Speech, Language, and Hearing Research*, vol. 26, no. 3, pp. 476–480, 1983.

[22] M. A. Redenbaugh and A. R. Reich, "Correspondence between an accelerometric nasal/voice amplitude ratio and listeners' direct magnitude estimations of hypernasality," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 2, pp. 273–281, 1985.

[23] D. C. Spriestersbach, "Assessing nasal quality in cleft palate speech of children," *Journal of Speech and Hearing Disorders*, vol. 20, no. 3, pp. 266–270, 1955.

[24] K. H. Keuning, G. H. Wieneke, H. A. Van Wijngaarden, and P. H. Dejonckere, "The correlation between nasalance and a differentiated perceptual rating of speech in Dutch patients with velopharyngeal insufficiency," *The Cleft Palate-Craniofacial Journal*, vol. 39, no. 3, pp. 277–284, 2002.

[25] M. K. Huffman and R. A. Krakow, *Nasals, Nasalization, and the Velum.* Academic Press, 1993, vol. 5.

[26] P. Vijayalakshmi, M. R. Reddy, and D. O'Shaughnessy, "Acoustic Analysis and Detection of Hypernasality Using a Group Delay Function," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 4, pp. 621–629, 2007.

[27] G. Fant, *Acoustic theory of speech production.* The Hague, Netherlands: Mouton, 1960.

[28] S. Hawkins and K. N. Stevens, "Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels," *The Journal of the Acoustical Society of America*, vol. 77, no. 4, pp. 1560–1574, Apr 1985.

[29] A. S. House and K. N. Stevens, "Analog studies of the nasalization of vowels," *Journal of Speech and Hearing Disorders*, vol. 21, no. 2, pp. 218–232, 1956.

[30] S. Maeda, "Acoustics of vowel nasalization and articulatory shifts in french nasal vowels," in *Nasals, Nasalization, and the Velum.* Elsevier, 1993, pp. 147–167.

[31] D. Cairns, J. H. Hansen, J. E. Riski *et al.*, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 1, pp. 35–45, 1996.

[32] A. Maier, F. Hönig, T. Bocklet, E. Nöth, F. Stelzle, E. Nkenke, and M. Schuster, "Automatic detection of articulation disorders in children with cleft lip and palate," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2589–2602, 2009.

[33] D. K. Rah, Y. I. Ko, C. Lee, and D. W. Kim, "A Noninvasive Estimation of Hypernasality Using a Linear Predictive Model," *Annals of Biomedical Engineering*, vol. 29, no. 7, pp. 587–594, 2001.

[34] J. R. Orozco-Arroyave, S. M. Rendón, A. M. Álvarez-Meza, J. D. Arias-Londoño, E. Delgado-Trejos, J. F. V. Bonilla, and C. G. Castellanos-Domínguez, "Automatic Selection of Acoustic and Non-Linear Dynamic Features in Voice Signals for Hypernasality Detection." in *Proc. Interspeech*, 2011, pp. 529–532.

[35] S. M. Rendón, J. O. Arroyave, J. V. Bonilla, J. A. Londoño, and C. C. Domínguez, "Automatic Detection of Hypernasality in Children," in *Proc. International Work-Conference on the Interplay Between Natural and Artificial Computation.* Springer, 2011, pp. 167–174.

[36] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla, and E. Nöth, "Automatic detection of hypernasal speech signals using nonlinear and entropy measurements." in *Proc. Interspeech*, 2012, pp. 2029–2032.

[37] G.-S. Lee, C.-P. Wang, C. C. Yang, and T. B. Kuo, "Voice Low Tone to High Tone Ratio: A Potential Quantitative Index For Vowel [a:] and Its Nasalization," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 7, pp. 1437–1439, 2006.

[38] L. He, J. Zhang, Q. Liu, H. Yin, and M. Lech, "Automatic evaluation of hypernasality and consonant misarticulation in cleft palate speech," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1298–1301, 2014.

[39] M. Golabbakhsh, F. Abnavi, M. Kadkhodaei Elyaderani, F. Derakhshandeh, F. Khanlar, P. Rong, and D. P. Kuehn, "Automatic identification of hypernasality in normal and cleft lip and palate patients with acoustic analysis of speech," *The Journal of the Acoustical Society of America*, vol. 141, no. 2, pp. 929–935, 2017.

[40] R. Kataoka, D. W. Warren, D. J. Zajac, R. Mayo, and R. W. Lutz, "The relationship between spectral characteristics and perceived hypernasality in children," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2181–2189, 2001.

[41] K. Nikitha, S. Kalita, C. M. Vikram, M. Pushpavathi, and S. R. M. Prasanna, "Hypernasality Severity Analysis in Cleft Lip and Palate Speech Using Vowel Space Area." in *Proc. Interspeech*, 2017, pp. 1829–1833.

[42] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 309–319, 1979.

[43] D. R. Dickson, "An acoustic study of nasality," *Journal of Speech and Hearing Research*, vol. 5, no. 2, pp. 103–111, 1962.

[44] Y. Bayya and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[45] R. Sinha and S. Ghai, "On the use of pitch normalization for improving children's speech recognition," in *Proc. Tenth Annual Conference of the International Speech Communication Association*, 2009.

[46] J. Chen, Y. Huang, Q. Li, and K. K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 258–261, 2004.

[47] J. Dang, K. Honda, and H. Suzuki, "Morphological and acoustical analysis of the nasal and the paranasal cavities," *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2088–2100, 1994.

[48] G. Bjuggren and G. Fant, "The nasal cavity structures," *STL-QPSR*, vol. 5, no. 4, pp. 5–7, 1964.

[49] P. Tarun, "Analysis, vocal-tract modeling and automatic detection of vowel nasalization," Ph.D. dissertation, PhD Thesis U. of Maryland, Tech. Rep, 2005.

[50] P. Delattre, "Les Attributs Acoustiques De La Na-Salité Vocalique Et Consonantique," *Studia linguistica*, vol. 8, no. 1-2, pp. 103–109, 1954.

[51] S. Hattori, K. Yamamoto, and O. Fujimura, "Nasalization of vowels in relation to nasals," *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 267–274, 1958.

[52] R. Carre and J. Tribolet, "Acoustic characteristics of vowel nasalization," *The Journal of the Acoustical Society of America*, vol. 55, no. S1, pp. S20–S21, 1974.

[53] M. F. Schwartz, "The acoustics of normal and nasal vowel production," *The Cleft Palate Journal*, vol. 5, no. 2, pp. 125–140, 1968.

[54] O. Fujimura and J. Lindqvist, "Sweep-tone measurements of vocal-tract characteristics," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 541–558, 1971.

[55] S. Maeda, "The role of the sinus cavities in the production of nasal vowels," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 7, 1982, pp. 911–914.

[56] J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3374–3383, 1996.

[57] J. Lindqvist-Gauffin and J. Sundberg, "Acoustic properties of the nasal tract," *Phonetica*, vol. 33, no. 3, pp. 161–168, 1976.

[58] K. N. Steves, "Some acoustical and perceptual correlates of nasal vowels," *Festschrift fur Ilse Lehiste*, pp. 241–254, 1987.

[59] E. Bognar and H. Fujisaki, "Analysis, synthesis and perception of the French nasal vowels," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 11, 1986, pp. 1601–1604.

[60] P. S. Beddor and S. Hawkins, "The influence of spectral prominence on perceived vowel quality," *The Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2684–2704, 1990.

[61] T. Arai, "Formant shift in nasalization of vowels," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2541–2541, 2004.

[62] M. Y. Chen, "Acoustic parameters of nasalized vowels in hearing impaired and normal hearing speakers," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2443–2453, Nov 1995.

[63] ——, "Acoustic correlates of nasality in speech," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.

[64] ——, "Acoustic correlates of English and French nasalized vowels," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2360–2370, 1997.

[65] W. Styler, "On the acoustical features of vowel nasality in English and French," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2469–2482, 2017.

[66] P. Milenkovic and F. Mo, "Glottal inverse filtering of nasalized vowels," *The Journal of the Acoustical Society of America*, vol. 80, no. S1, pp. S19–S19, 1986.

[67] G. Feng and E. Castelli, "Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3694–3706, 1996.

[68] D. Nguyen and B. Guérin, "Effects of nasal coupling on the vowels," *The Journal of the Acoustical Society of America*, vol. 67, no. S1, pp. S94–S94, 1980.

[69] T. Pruthi, C. Y. Espy-Wilson, and B. H. Story, "Simulation and analysis of nasalized vowels based on magnetic resonance imaging data," *The Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3858–3873, 2007.

[70] J. Glass and V. Zue, "Detection of nasalized vowels in American English," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 10, 1985, pp. 1569–1572.

[71] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for the automatic detection of vowel nasalization," in *Proc. Eighth Annual Conference of the International Speech Communication Association*, 2007.

[72] S. Najnin and C. Shahnaz, "A detection and classification method for nasalized vowels in noise using product spectrum based cepstra," *International Journal of Speech Technology*, vol. 18, no. 1, pp. 97–111, 2015.

[73] S. Ha and D. P. Kuehn, "Temporal characteristics of nasalization in speakers with and without cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 48, no. 2, pp. 134–144, 2011.

[74] Y. Kozaki-Yamaguchi, N. Suzuki, Y. Fujita, H. Yoshimasu, M. Akagi, and T. Amagasa, "Perception of hypernasality and its physical correlates," *Oral Science International*, vol. 2, no. 1, pp. 21–35, 2005.

[75] M. Eshghi, M. M. Alemi, and M. Eshghi, "Vowel nasalization might affect the envelop of the vowel signal by reducing the magnitude of the rising and falling slope amplitude," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2304–2304, 2015.

[76] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, 2002.

[77] C. Vikram, N. Adiga, and S. R. M. Prasanna, "Spectral Enhancement of Cleft Lip and Palate Speech." in *Proc. Interspeech*, 2016, pp. 117–121.

[78] D. A. Cairns, J. H. Hansen, and J. F. Kaiser, "Recent advances in hypernasal speech detection using the nonlinear Teager energy operator," in *Proc. IEEE Fourth International Conference on Spoken Language Processing (ICSLP)*, vol. 2, 1996, pp. 780–783.

[79] J. F. Kaiser, "On a simple algorithm to calculate the'energy'of a signal," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1990, pp. 381–384.

[80] C. De La Cruz, "Hypernasal Speech Analysis via Emperical Mode Decomposition and the Teager-Kasiser Energy Operator," 2016.

[81] G.-S. Lee, C.-P. Wang, and S. Fu, "Evaluation of Hypernasality in Vowels using Voice Low Tone to High Tone Ratio," *The Cleft Palate-Craniofacial Journal*, vol. 46, no. 1, pp. 47–52, 2009.

[82] Y. Huang, C. Yan, and Q. Xu, "On the difference between empirical mode decomposition and Hilbert vibration decomposition for earthquake motion records," in *Proc. 15th World Conference on Earthquake Engineering*, 2012.

[83] P. Vijayalakshmi and M. RamasubbaReddy, "The analysis on band-limited hypernasal speech using group delay based formant extraction technique," in *Proc. Eurospeech*, 2005.

[84] H. Murthy, V. Gadde *et al.*, "The modified group delay function and its application to phoneme recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, vol. 1, 2003, pp. 68–71.

[85] P. Vijayalakshmi, T. Nagarajan, and J. Rav, "Selective pole modification-based technique for the analysis and detection of hypernasality," in *IEEE Region 10 Conference (TENCON)*, 2009, pp. 1–5.

[86] P. Vijayalakshmi and M. RamasubbaReddy, "Detection of hypernasality using statistical pattern classifiers," in *Proc. Eurospeech*, 2005.

[87] J. Qian, F. Fu, X. Liu, L. He, H. Yin, and H. Zhang, "The analysis and detection of hypernasality based on a formant extraction algorithm," in *Journal of Physics: Conference Series*, vol. 887, no. 1. IOP Publishing, 2017, p. 012082.

[88] E. Akafi, M. Vali, and N. Moradi, "Detection of hypernasal speech in children with cleft palate," in *Proc. IEEE 19th Iranian Conference of Biomedical Engineering (ICBME)*, 2012, pp. 237–241.

[89] S. Ghai and R. Sinha, "Exploring the role of spectral smoothing in context of children's speech recognition," in *Proc. Eurospeech*, 2009.

[90] X. Wang, M. Tang, S. Yang, H. Yin, H. Huang, and L. He, "Automatic Hypernasality Detection in Cleft Palate Speech Using CNN," *Circuits, Systems, and Signal Processing*, pp. 1–27, 2019.

[91] Y. Liu, X. Wang, Y. Hang, L. He, H. Yin, and C. Liu, "Hypemasality detection in cleft palate speech based on natural computation," in *Proc. IEEE 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2016, pp. 523–528.

[92] C. Vikram, A. Tripathi, S. Kalita, and S. R. M. Prasanna, "Estimation of Hypernasality Scores from Cleft Lip and Palate Speech." in *Proc. Interspeech*, 2018, pp. 1701–1705.

[93] J. Zhang, S. Yang, X. Wang, M. Tang, H. Yin, and L. He, "Automatic hypernasality grade assessment in cleft palate speech based on the spectral envelope method," *Biomedical Engineering/Biomedizinische Technik*, 2019.

[94] A. P. Vogel, H. M. Ibrahim, S. Reilly, and N. Kilpatrick, "A comparative study of two acoustic measures of hypernasality," *Journal of Speech, Language, and Hearing Research*, 2009.

[95] S. Berkowitz, "Diagnostic procedures and instruments used in the assessment and treatment of speech," in *Cleft Lip and Palate*. Springer, 2006, pp. 615–620.

[96] D. A. Stringer and M. Witzel, "Velopharyngeal insufficiency on videofluoroscopy: Comparison of projections," *American Journal of Roentgenology*, vol. 146, no. 1, pp. 15–19, 1986.

[97] A. J. Beer, P. Hellerhoff, A. Zimmermann, K. Mady, R. Sader, E. J. Rummeny, and C. Hannig, "Dynamic near-real-time magnetic resonance imaging for analyzing the velopharyngeal closure in comparison with videofluoroscopy," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 20, no. 5, pp. 791–797, 2004.

[98] C. Drissi, M. Mitrofanoff, C. Talandier, C. Falip, V. Le Couls, and C. Adamsbaum, "Feasibility of dynamic MRI for evaluating velopharyngeal insufficiency in children," *European Radiology*, vol. 21, no. 7, pp. 1462–1469, 2011.

[99] S. Vadodaria, T. Goodacre, and P. Anslow, "Does MRI contribute to the investigation of palatal function?" *British Journal of Plastic Surgery*, vol. 53, no. 3, pp. 191–199, 2000.

[100] M. R. Rowe and L. L. D'Antonio, "Velopharyngeal dysfunction: evolving developments in evaluation," *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 13, no. 6, pp. 366–370, 2005.

[101] S. Maturo, A. Silver, K. Nimkin, P. Sagar, J. Ashland, A. J. Van Der Kouwe, and C. Hartnick, "Mri with synchronized audio to evaluate velopharyngeal insufficiency," 2012.

[102] J. L. Perry, D. P. Kuehn, J. M. Wachtel, J. S. Bailey, and L. L. Luginbuhl, "Using magnetic resonance imaging for early assessment of submucous cleft palate: A case report," *The Cleft Palate-Craniofacial Journal*, vol. 49, no. 4, pp. 35–41, 2012.

[103] M. P. Karnell, "Instrumental assessment of velopharyngeal closure for speech," in *Seminars in Speech and Language*, vol. 32, no. 02. © Thieme Medical Publishers, 2011, pp. 168–178.

[104] D. P. Kuehn and K. T. Moller, "Speech and language issues in the cleft palate population: the state of the art," *The Cleft Palate-Craniofacial Journal*, vol. 37, no. 4, pp. 1–35, 2000.

[105] I. Honjo, T. Mitoma, K. Ushiro, and M. Kawano, "Evaluation of velopharyngeal closure by ct scan and endoscopy." *Plastic and Reconstructive Surgery*, vol. 74, no. 5, pp. 620–627, 1984.

[106] S. Suri, A. Utreja, N. Khandelwal, and S. K. Mago, "Craniofacial computerized tomography analysis of the midface of patients with repaired complete unilateral cleft lip and palate," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 134, no. 3, pp. 418–429, 2008.

[107] C. Hawkins, "Evaluation of a real-time ultrasound scanner in assessing lateral pharnygeal wall motion during speech," *Cleft Palate*, vol. 15, pp. 161–166, 1978.

[108] K. M. Van Lierde, M. De Bodt, I. Baetens, V. Schrauwen, and P. Van Cauwenberge, "Outcome of treatment regarding articulation, resonance and voice in Flemish adults with unilateral and bilateral cleft palate," *Folia Phoniatrica et Logopaedica*, vol. 55, no. 2, pp. 80–90, 2003.

[109] V. D. de Pochat, N. Alonso, R. R. da Silva Mendes, P. R. Gravina, E. V. Cronenberg, and J. V. L. Meneses, "Assessment of nasal patency after rhinoplasty through the Glatzel mirror," *International Archives of Otorhinolaryngology*, vol. 16, no. 03, pp. 341–345, 2012.

[110] A. Main, S. Kelly, and G. Manley, "Notes and discussion instrumental assessment and treatment of hypernasality, following maxillofacial surgery, using snors: a single case study," *International Journal of Language & Communication Disorders*, vol. 34, no. 2, pp. 223–238, 1999.

[111] J. Karling, O. Larson, R. Leanderson, K. Galyas, and A. De Serpa-Leitâo, "Noram–An Instrument Used in the Assessment of Hypernasality: A Clinical Investigation," *The Cleft Palate-Craniofacial Journal*, vol. 30, no. 2, pp. 135–140, 1993.

[112] R. J. Baken and R. F. Orlikoff, *Clinical measurement of speech and voice*. Cengage Learning, 2000.

[113] A. C. Nichols, "Nasalance statistics for two Mexican populations," *The Cleft Palate-Craniofacial Journal*, vol. 36, no. 1, pp. 57–63, 1999.

[114] P. Roche, "Characteristics of nasalance in speakers of western Canadian English and French," *Journal of Speech Language Pathology and Audiology*, vol. 22, pp. 94–103, 1998.

[115] K. Brunnegård and J. van Doorn, "Normative data on nasalance scores for Swedish as measured on the Nasometer: Influence of dialect, gender, and age," *Clinical Linguistics & Phonetics*, vol. 23, no. 1, pp. 58–69, 2009.

[116] R. Mayo, L. A. Floyd, D. W. Warren, R. M. Dalston, and C. M. Mayo, "Nasalance and nasal area values: cross-racial study," *The Cleft Palate-Craniofacial Journal*, vol. 33, no. 2, pp. 143–149, 1996.

[117] AIISH, "All India Institute of speech and Hearing, Mysore, India." [Online]. Available: web-site:http://www.aiishmysore.in

[118] K. Sjölander and J. Beskow, "Wavesurfer-an open source speech tool," in *Proc. Sixth International Conference on Spoken Language Processing*, 2000.

[119] K. N. Stevens, *Acoustic phonetics, book,*. MIT press, 1998.

[120] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[121] B. Sharma and S. R. M. Prasanna, "Sonority measurement using system, source, and suprasegmental information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 505–518, 2016.

[122] B. D. Sarma and S. R. M. Prasanna, "Analysis of Vocal Tract Constrictions using Zero Frequency Filtering," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1481–1485, 2014.

[123] K. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[124] S. R. Kadiri and B. Yegnanarayana, "Analysis of singing voice for epoch extraction using Zero Frequency Filtering method," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4260–4264.

[125] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[126] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The Journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[127] G. D. Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

[128] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1329–1334, 1986.

[129] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio–a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[130] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Physica D: Nonlinear Phenomena*, vol. 9, no. 1-2, pp. 189–208, 1983.

[131] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, "A practical method for calculating largest Lyapunov exponents from small data sets," *Physica D: Nonlinear Phenomena*, vol. 65, no. 1-2, pp. 117–134, 1993.

[132] M. Aboy, R. Hornero, D. Abásolo, and D. Álvarez, "Interpretation of the Lempel-Ziv complexity measure in the context of biomedical signal analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 11, pp. 2282–2288, 2006.

[133] H. Hurst, R. Black, and Y. Simaika, "Long-term storage: an experimental study Constable," *London, UK*, 1965.

[134] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of biological signals," *Physical Review E*, vol. 71, no. 2, p. 021906, 2005.

[135] J. S. Richman, D. E. Lake, and J. R. Moorman, "Sample entropy," in *Methods in Enzymology*. Elsevier, 2004, vol. 384, pp. 172–184.

[136] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomedical Engineering Online*, vol. 6, no. 1, p. 23, 2007.

[137] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge university press, 2000.

[138] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[139] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[140] M. Novotny, J. Rusz, R. Cmejla, and E. Ruzicka, "Automatic evaluation of articulatory disorders in Parkinson's disease," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 9, pp. 1366–1378, 2014.

[141] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[142] S. Ramamohan and S. Dandapat, "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 737–746, 2006.

[143] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[144] H. Hermansky, H. Fujisaki, and Y. Sato, "Spectral envelope sampling and interpolation in linear predictive analysis of speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 9, 1984, pp. 53–56.

[145] J. Pohjalainen, O. Räsänen, and S. Kadioglu, "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits," *Computer Speech & Language*, vol. 29, no. 1, pp. 145–171, 2015.

[146] A. Al-Ani and M. Deriche, "Feature selection using a mutual information based measure," in *Object Recognition Supported by User Interaction for Service Robots*, vol. 4. IEEE, 2002, pp. 82–85.

[147] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. Eurospeech*, 2003.

[148] M. Anand Joseph, S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. Interspeech*, 2006.

[149] B. Yegnanarayana, D. Saikia, and T. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 3, pp. 610–623, Jun 1984.

[150] B. Yegnanarayana and H. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, 1992.

[151] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. Prentice hall, 1993.

[152] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1638–1640, 1978.

[153] K. R. Rao and P. Yip, *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014.

[154] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-Adaptive Front-End Features for Robust Children's ASR." in *Proc. Interspeech*, 2016, pp. 3459–3463.

[155] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Spectral Moment Features Augmented by Low Order Cepstral Coefficients For Robust ASR," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 551–554, 2010.

[156] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3795–3806, 1996.

[157] S. Scherer, L.-P. Morency, J. Gratch, and J. Pestian, "Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4789–4793.

[158] E. A. Wieland, E. B. Burnham, M. Kondaurova, T. R. Bergeson, and L. C. Dilley, "Vowel space characteristics of speech directed to children with and without hearing loss," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 2, pp. 254–267, 2015.

[159] R. A. Fox and E. Jacewicz, "Reconceptualizing the vowel space in analyzing regional dialect variation and sound change in American English," *The Journal of the Acoustical Society of America*, vol. 142, no. 1, pp. 444–459, 2017.

[160] N. Adiga and S. R. M. Prasanna, "Detection of Glottal Activity Using Different Attributes of Source Information," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 2107–2111, 2015.

List of Publications

Journal Publications

- **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Detection and assessment of hypernasality in repaired cleft palate speech using vocal tract and residual features," *J. Acoust. Soc. Am.* 146(6), 4211-4223(2019).
- **A. K. Dubey**, A. Tripathi, S. R. M. Prasanna, and S. Dandapat, "Detection of hypernasality based on vowel space area," *J. Acoust. Soc. Am.* 143(5), EL412-EL417(2018).
- **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Sinusoidal model-based hypernasality detection in cleft palate speech using CVCV sequence," *Speech Communication*(2020).

Conference and Workshop Publications

1. **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat,"Pitch-Adaptive Front-end Feature for Hypernasality detection," in *Proc. Interspeech*, 372-376(2018).
2. **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Hypernasality detection using Zero time windowing," in IEEE *International Conference on Signal Processing and Communications (SPCOM)*, 105-109(2018).
3. **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Zero time windowing analysis of hypernasality in speech of cleft lip and palate children," in IEEE *Twenty Second National Conference on Communications (NCC)*, 1-6(2016).

Other Publications

1. Sishir Kalita, **A. K. Dubey**, C. M. Vikram, Protima Nomo Sudra, S. R. M. Prasanna, and S. Dandapat, Excitation source based analysis of cleft lip and palate speech, under major revision in Speech Communication journal.
2. **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Hypernasality Severity detection using Constant Q Cepstral Coefficients." in *Proc. Interspeech*, 4554-4558(2019).
3. **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Automatic detection of hypernasality using modified group delay feature," in *Proc. Workshop on Speech Processing for Voice , Speech and Hearing Disorders (WSPD)*, Mysure, India, September 2018.

4. **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Sinusoidal model based features for analysis of hypernasality severity," in *Proc. Workshop on Speech Processing for Voice , Speech and Hearing Disorders (WSPD)*, Mysure, India, September 2018.

5. **A. K. Dubey**, S. R. M. Prasanna, and S. Dandapat, "Zero time windowing based severity analysis of hypernasal speech," in IEEE *Region 10 Conference (TENCON)*, 970-974(2016).