Enhancement of Cleft Lip and Palate Speech

Protima Nomo Sudro

Enhancement of Cleft Lip and Palate Speech

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

Protima Nomo Sudro

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

September 2021

Certificate

This is to certify that the thesis entitled "**Enhancement of Cleft Lip and Palate speech**", submitted by **PROTIMA NOMO SUDRO** (146102035), 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 her 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:
Dharwad.

Prof. S. R. Mahadeva Prasanna
Professor
Dept. of Electrical Engg.
Indian Institute of Technology Dharwad
Dharwad - 580 011, Karnataka, India.

Dated:
Guwahati.

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

To

Maa and Baba

for their love, support, and blessing

&

My sister and brothers

for their encouragement

Acknowledgements

I express my deep and sincere gratitude to my research supervisors, Prof. S. R. M. Prasanna and Prof. Rohit Sinha 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. S. Dandapat, the Chairman of the Doctoral Committee 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.

This thesis would not become possible without the help and support of Speech-language pathologists and staff of All India Institute of Speech and Hearing, 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.

My sincere thanks goes 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 close friends Dr. Vikram CM, Dr. Sishir Kalita, and Dr. Akhilesh Dubey, 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, Shikha, Himakshi, Sarfaraz, Mrinmoy, Sandeep, Saswati, Balaji, Sreeram, Sukanya, Vineeta, Prabhakar, Shoubhik, Deepika, Anik, Brij, Alex, Ato, Mousumi, Sushmita, Salman, and Rohan for their direct/indirect contributions during my stay at IIT Guwahati. 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 my friends Sumi, Kamakshi, Gayatri, Trusna, Uddipana, Kasturi, Gautam, Tarique, Wendy, Aparajita, Tanushree, Niharika, Abhishek for being there with me all the time. I would like to thank my parents, sister, and brothers for their their blessings, support, encouragement and prayers throughout my Ph.D. Without their support, it would not have been possible to come this far. I would also like to thank Mr. Dey for the support and encouragement, a late joinee in the list but an important one. Finally, I thank Shri Krishna for all the blessings and bearing the confrontations I made during my hard times.

Protima Nomo Sudro

Abstract

The individuals with cleft lip and palate (CLP) suffer from speech disorders due to articulatory impairments. As a result, measurable reductions are observed in the speech intelligibility and quality. Despite the advances in surgical management, the problems of articulation and resonance still remain to some degree in the individuals with CLP. For repaired CLP speakers, another direction of improvement in the speech intelligibility could be achieved through exploration of signal processing algorithms. Among all the speech disorders demonstrated by individuals with CLP, certain disorders relatively have a more severe impact on speech. In CLP speech, it is reported that the speech intelligibility is affected by two factors mainly: hypernasality and articulation errors. These two speech disorders are addressed in this thesis. In this thesis, we consider enhancement of CLP speech intelligibility by modifying frequently observed phoneme-specific distortions (fricative misarticulation, stop misarticulation, and vowel nasalization). The CLP speech modification is performed with an assumption that the ground truth of the phoneme distortions is available.

The first work in the thesis addresses the modification of fricative /s/ misarticulation. Based on the deviant characteristics, the misarticulated fricatives are segmented automatically followed by categorization of the type of distortion into palatalization, phoneme specific nasal air emission, and glottal stop. The fricative distortions that involve change in place of articulation are corrected using spectral compression and spectral tilt modification. While the insertion method is used when both change in the place and manner of articulation are observed.

In the next work, misarticulated stops are modified. As the stop consonants play an important role in speech perceptivity, it is important to address them as well. The work focuses on three unvoiced stop consonants /k/, /t/, and /T/. Three type of misarticulations are studied: glottal, palatal, and velar stop substitutions. An event-based modification approach is used to correct the misarticulated stops, where at first, automatic detection of burst onset and vowel onset events is carried out. The misarticulated stops are modified using spectral conversion method.

For an entire word enhancement, apart from stops, the vowel distortion needs to be addressed. The nasalized vowels (hypernasality) are observed to influence the speech intelligibility and quality both. Hence, vowel modification is performed as the third work. The issues related to nasalization is studied for vowels /a/, /i/, and /u/. In this work, the CLP distortions are analyzed in children speech and they exhibit high-pitch speech. Additionally, hypernasality introduces nasal resonances in the oral sounds,

with consistent nasal resonances in the low-frequency region. Therefore, an accurate representation of the spectral envelope is necessary, for which extended weighted linear prediction (XLP) method is used. The transformation is achieved using spectral conversion method. The deviated spectral characteristics results in interference/additional signal components in the residual signal. Therefore, a weighting function is used for de-emphasizing the interfering signal components in the XLP residual signal.

Finally the word-level intelligibility is attempted by combining the specific phoneme modification techniques discussed in the earlier works. Several issues exist in performing the entire word-level intelligibility because many times, both articulation error and hypernasality are observed in the same word. It is challenging to detect such misarticulations in an unsupervised method. Hence, with certain assumptions and prior knowledge, the enhancement task is carried out.

To transform the CLP speech, different attempts are made in the thesis. The notable contributions of the thesis are listed below:

- A database is developed for analysis and enhancing the CLP speech. Database comprised of nonsensical words, vowel phonations, meaningful words, and short phrases. However, only some nonsensical words and vowel phonations are used in this thesis.
- Misarticulated fricative /s/ is first studied because it is observed as one of the frequently occurring speech distortions in the database and findings of various studies also support the same.
- Misarticulated unvoiced stops /k/, /t/, and /T/ are analyzed and modified.
- Nasalized vowels /a/, /i/, and /u/ are modified using temporal as well as spectral processing.
- Finally, phoneme specific modification techniques are combined to achieve entire word-level intelligibility.

Keywords: cleft lip and palate speech, articulation error, hypernasality, intelligibility, speech enhancement.

Contents

List of Figures

List of Tables

List of Acronyms

AIISH	All India Institute of Speech and Hearing
ASR	Automatic speech recognition
B	Burst
BER	Band energy ratio
BT	Burst plus transition region
CLP	Cleft lip and palate
CV	Consonant-vowel
CVCV	Consonant-vowel-consonant-vowel
CycleGAN	Cycle consistent generative adversarial network
DNN	Deep neural network
DSC	Dominant spectral centroid
DRNN	Deep recurrent neural network
EL	Electrolarynx
ED	Euclidean distance
EW	Entire word
F0	Fundamental frequency
F1	First formant
FFT	Fast Fourier transform
FCN	Fully convolutional neural network
Fs	Sampling frequency
FVFV	Fricative-vowel-fricative-vowel
FV	Fricative-vowel
GAN	Generative adversarial network
GCI	Glottal closure instant

GMM	Gaussian mixture modeling
GS	Glottal stop
HLR	High-to-low frequency energy ratio
HMM	Hidden markov model
HN	Hypernasal
ICC	Intra-class correlation coefficient
LP	Linear prediction
LSTM	Long short-term memory networks
M1	Spectral centroid/ first spectral moment
MAP	Maximum *a posteriori*
MCD	Mel-cepstral distortion
MFCC	Mel-frequency cepstral coefficient
ML	Maximum likelihood
MLP	Multilayer perceptron
MMSE	Minimum mean square error
MNSS	Maximum normalized spectral slope
MoA	Manner of articulation
MOS	Mean opinion score
MWIP	Maximum weighted inner product
NMF	Nonnegative matrix factorization
NS	Normal fricative /s/
PA	Palatalized articulation
PCC	Percentage of consonants correct
PI	Plosion index
PSNAE	Phoneme specific nasal air emission
PSD	Power spectral density
PSR	Peak-to-sidelobe ratio
P-STOI	Pathological short-time objective intelligibility
P-ESTOI	Pathological extended short-time objective intelligibility
PoA	Place of articulation

SoE	Strength of excitation
SSD	Speech sound disorder
SLP	Speech language pathologist
STE	Short-time energy
STSA	Short-time spectral amplitude
STRAIGHT	Speech transformation and representation using adaptive interpolation of weighted spectrum
SNR	Signal-to-noise ratio
SPSS	Statistical parametric speech synthesis
SVM	Support vector machine
TM	XLP residual modification
VTSM	Vocal tract system characteristics modification
TTS	Text-to-speech synthesis
VC	Voice conversion
VCV	Vowel-consonant-vowel
VLHR	Voice low tone to high tone ratio
VOP	Voice onset point
VOT	Voice onset time
VPD	Velopharyngeal dysfunction
VAWGAN	Variational autoencoding Wassertein generative adversarial network
WDA	Weighted denoising encoder
XLPC	Extended weighted linear prediction coefficient
XLPCC	Extended weighted linear prediction cepstral coefficient
XCD	XLPC cepstral distortion
ZFF	Zero frequency filtering
ZFFS	Zero frequency filtered signal

1

Introduction

Contents

Overview

Speech is a natural and convenient means of communication. It is used to express needs, emotions and share knowledge. A majority of people use speech effectively to communicate in a real-world environment. Regardless of the communication environment (presence of noise, reverberation), people understand how to adjust to deliver their message successfully and partake in the conversation with a wide range of communication partners. However, a group of people with speech sound disorders (SSD) such as dysarthria, cleft lip and palate (CLP), hearing impairment, laryngectomy, and glossectomy face difficulty in using speech effectively for communication. The problem varies across a variety of environments. Unfamiliar listeners find it hard to be involved in the conversation with most of the people who exhibit SSD. Clinically, the treatment of SSD requires surgical correction, prosthesis, and speech therapy. Besides clinical intervention, studies report the modification of SSD based on signal processing techniques. In a similar direction, this work aims at improving the speech intelligibility and quality of one of the SSD: CLP speech. The work attempts to examine the intelligibility and quality differences between CLP and healthy (non-CLP) speech. The study analyzed the factors impacting the CLP speech caused by vowel nasalization and specific obstruent errors such as misarticulated fricative and stop consonants. The present study deals with phoneme-specific modifications. The deviant acoustic characteristics that contribute to the intelligibility and quality distortion are identified and transformed. The modified speech segments replaced the distorted segments in the original speech to observe the impact of the enhancement method. Further, all the specific phoneme modification methods are used to improve entire word-level intelligibility and quality.

1.1 Cleft lip and palate speech

Cleft lip and palate (CLP) is one of the common congenital disorder of the craniofacial region. Besides other problems associated with cleft and craniofacial impairment, it mostly impacts the articulatory structures. Impaired articulatory systems lead to difficulty in speech production. Therefore, individuals with clefts are at risk for delays in acquiring speech skills and language development. Individuals with CLP require surgical repair (such as, maxilofacial surgery, palatal surgery), prosthesis, and speech therapy to establish appropriate oral-motor skills. However, speech disorders may persist due to velopharyngeal dysfunction (VPD), oronasal fistula, and mislearning even after clinical intervention [2]. The presence of VPD results in the loss of oral pressure during the production of

pressure-sensitive obstruents and results in the misarticulation of obstruents (i.e., stops, fricatives, and affricates). VPD also results in hypernasality. In the case of an oronasal fistula, air escapes through the fistula opening. To avoid the airflow through the fistula, an individual made a constriction behind the fistula and end up in producing deviated speech sound. An oronasal fistula may result in weak or omitted consonants and the nasalization of consonants. Mislearning is an articulation disorder that results in the substitution of certain nasal or pharyngeal sounds for oral sounds [3]. Based on the nature of speech distortion, the different type of speech disorders demonstrated by the CLP individuals are categorized into: hypernasality, hyponasality, misarticulations, nasal air emission, and voice disorders.

Hypernasality corresponds to a resonance disorder, and the presence of nasal resonances during speech production has an excessively perceptible nasal quality [4]. Mostly, the voiced sounds are nasalized, and the nasal consonants tend to replace the obstruents due to severe hypernasality. All these factors affect speech intelligibility and quality both [2, 5, 6]. Hyponasality is a type of abnormal resonance that occurs due to blockage in the nasal cavity entrance and hence, affects the production of nasal consonants. Besides hypernasality and hyponasality, the CLP speech intelligibility is also affected by misarticulations (articulation error) such as weak consonant or nasalized consonant or glottal stop or pharyngeal substitution or velar substitution [2, 3]. Misarticulations are produced either due to the structural or functional disorder or both. Misarticulations in CLP speech are categorized into two types: obligatory and compensatory errors [7]. Obligatory distortions occur when articulation is normal, but the structure is abnormal. Weak and nasalized consonants are examples of obligatory errors. Compensatory errors are those that occur when articulation placement is changed to avoid the abnormal structure. Compensatory errors include glottal, pharyngeal, and palatal substitutions. The glottal and pharyngeal substitutions produced for the oral targets are due to air escape through the nasal cavity. Nasal air emission (NAE) consists of a turbulence noise source created in the nasal cavity. The turbulence noise produced in the nasal cavity is exhaled forcibly, which becomes a part of the generated speech signal. NAE also influences the speech intelligibility [2]. Voice disorder results in an alteration in the normal phonatory quality of the voice. It is characterized by breathiness, hoarseness, low intensity, and glottal fry [3]. Voice disorders may or may not impact speech intelligibility.

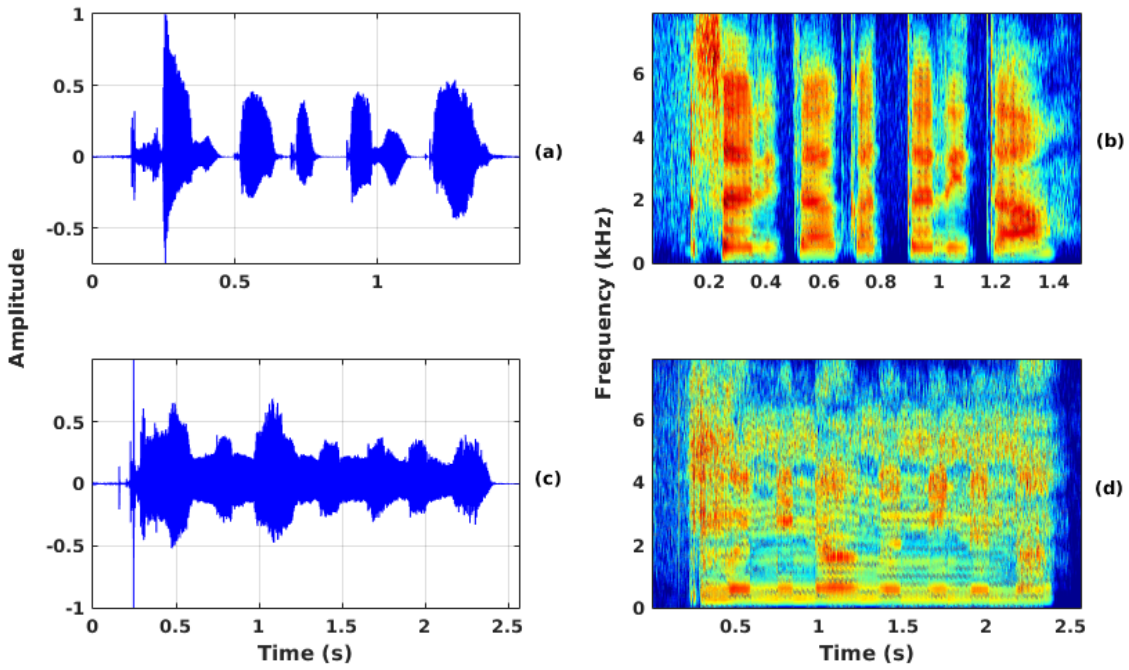

Figure 1.1: Illustration of the waveform and spectrogram of the sentence /Sarita kattari ta/ for (a)-(b) non-CLP speaker and (c)-(d) CLP speaker.

1.1.1 Primary factors associated with CLP speech distortion

Among the many speech disorders demonstrated by individuals with CLP, certain disorders relatively have a more severe impact on speech. It is reported that in the case of individuals with CLP, speech intelligibility is primarily affected by hypernasality and articulation errors [8]. Hence, in this subsection, a brief review of the impact of articulation error and hypernasality in CLP speech is presented to draw insights for the study.

For an illustration, the impact of articulation error and hypernasality is depicted in Figure 1.1. From Figure 1.1, the differences between non-CLP (Figure 1.1 (a)-(b)) and CLP (Figure 1.1 (c)-(d)) speech can be observed in terms of burst evidence, formant transitions, spectral energy in the low-frequency region of the vowels, spectral energy in the high-frequency region of obstruents, and pauses between words. From Figure 1.1, it is also observed that the phonemes are not easily distinguishable in CLP speech. The utterance showed in Figure 1.1 (c)-(d) is considered from a CLP speaker with moderate-severe speech intelligibility distortion. Due to severity, both articulation error and nasalization are observed. The acoustic characteristics of the intended speech sounds are distorted because of

the unintentional production of nasalized vowels and nasal cognates, as certain voiced stops share the same place of articulation (PoA) with nasal consonants [8].

Several studies in the literature have also observed that the speech disorders shown in Figure 1.1 reduces the CLP speech intelligibility. A study in [9] analyzed the acoustic deviations of CLP speech using the features extracted from the regions around the vowel onset points. The presence of nasal sound characteristics in the speech signals is also analyzed to determine the nasalization impact on speech intelligibility [10]. In the study, articulation and hypernasality scores are mapped to a measure using a trained regression model. The mapped measure is considered as the composite measure of intelligibility. The predicted intelligibility scores showed that the essential cues responsible for speech intelligibility are distorted in the case of CLP speech [3]. Based on the reported speech outcome measures, hypernasality is observed to reduce speech intelligibility or quality or both [11]. It is found that with an increase in hypernasality, the intelligibility decreases, and the authors suggested that hypernasality interacts with intelligibility [12]. Children with hypernasal speech are perceived by peers to be less pleasant, attractive, and intelligent [13–15]. As the ratings of nasality increases, the social acceptance ratings became more negative [16]. Another study examined the speech outcome of 110 adolescents with CLP, and they reported that 45% of the children demonstrated speech that was considered non-acceptable [3]. A study also said that articulation, nasal resonance, and nasal escape have a direct influence on the degree of intelligibility [17–20].

In a different study, a single-word intelligibility test was performed and found a significant correlation (r = 0.79) between perceptual intelligibility ratings and percentage of consonants correct (PCC) [21]. Articulation patterns and intelligibility were evaluated in 54 Vietnamese individuals with un-operated and operated cleft palate in the age range of $3-24$ years [22]. The authors indicate that there is relationship between the type of oral clefts and the proficiency of the speech skill attained by the speaker. Another study reported that PCC and intelligibility were positively correlated. In [23], the authors reported a significant correlation between articulation ratings and intelligibility ratings. A significant correlation (r = 0.715) is also reported between articulation errors and speech intelligibility in [24]. Additionally, a study investigated the relative contribution of speech disorders, namely articulation error, hypernasality, and voice disorder on CLP speech intelligibility [25]. From the study, it is observed that articulation error have the highest correlation $r = 0.919$ with the intelligibility ratings. Hypernasality exhibit a relatively lower correlation $r = 0.726$ and voice disorder have the least corre-

lation $r = 0.244$ with the intelligibility. This renders that CLP speech can be degraded by articulation error and hypernasality, making CLP speakers socially less confident while conversing with others. Since, the articulatory impairment affects the CLP speech intelligibility, it is given due focus. In the clinical settings the interdisciplinary management team (surgeons, dentists, speech therapists) try to improve the CLP speech intelligibility by using surgical intervention, prosthetics, and speech therapy.

1.2 Speech enhancement in clinical settings

Individuals with disordered speech require different types of interventions to establish normal speech production. Those who exhibit articulation and resonance problems may require both surgical management and speech therapy. On some occasions, an individual is not considered suitable for physical management, then prosthetic intervention is recommended. Most of the individuals with a history of CLP undergo speech therapy to minimize speech sound production difficulties caused by VPD and other reasons. Post-operatively, speech language pathologists (SLP) recommend speech therapy because it is essential to make the speaker learn the use of surgically corrected structure to produce speech like that of non-CLP [8, 26]. Several times, the surgical intervention may or may not result in functional correction of CLP speech, and deviant speech persists even after surgery. Therefore, speech therapy is recommended to obtain correct speech sound production.

Before proceeding with speech therapy, the SLPs evaluate the CLP speech using perceptual and instrumental measures. The SLP investigates the speech characteristics by engaging the individual with CLP in a conversation and make judgments regarding their intelligibility, articulation, voice, and resonance. The instrumental measures involve the techniques like nasometry, videofluoroscopy, nasopharyngoscopy, and electropalatography, to obtain crucial additional information [2]. The primary task in assessment is to determine whether the individual is stimulable or not for eliminating the compensatory placement errors with adequate correction of articulation placement. An SLP examines the relationship between intelligibility and other speech-language variables to select appropriate treatment methods. In some instances, different forms of supplemental evaluation procedures (such as visual, tactile, and auditory testing) are employed to more clearly determine the speech characteristics and its underlying causes [2, 3].

The individuals with CLP demonstrate a wide range of speech production problems. The SLP selects an appropriate strategy to enhance articulation or phonological development or general expres-

sive language functioning based on the type of impairment. At first, the SLP performs an auditory discrimination test, where digital recording equipment is used to help the individual discriminate against the correct and disordered speech. If the speaker with CLP cannot distinguish the contrast, the SLP further investigates other associated issues. A finding that the perception of the misarticulated speech sound is inadequate motivates the SLPs to include the perceptual component as a target in clinical intervention. The SLPs also examine whether an individual display problems on higher-level knowledge of phonemic categories using speech-production-perception task. The task assesses an individual ability to appreciate the phonological contrast between clear productions of the target sound and the substituted sound [27, 28]. Once the problem is identified, the SLPs considers the individual for speech therapy. In the case of articulation problems, the SLPs try to establish appropriate articulatory placement. If necessary, the SLPs work on specific phonemes in isolation, where they ask the individual to prolong phoneme placement. During speech therapy, biofeedback (visual, tactile, and auditory) is considered an effective treatment form. Although all forms of biofeedback were reported to be helpful, the auditory feedback is usually considered most effective as it entails the cognitive ability of an individual. In auditory feedback, the SLPs simulate the disordered speech sounds and present them along with the correct speech sounds. Through auditory discrimination and training an individual can achieve good understanding on acoustic quality of the speech. The SLPs pointed out that the correction of abnormal speech sounds requires motor learning and motor memory. Motor learning occurs through audio feedback while motor memory is acquired through many repetitions (practice) of the correct speech sounds. Hence, intensive speech therapy is considered an effective form of treatment for speech-language problems.

1.3 Issues associated with CLP speech distortion

The development and innovation in technology have eased our daily lifestyles. Specifically, speech-based applications such as automatic speech recognition (ASR), language identification have changed the way people interact with their devices [29]. Various potential applications of ASR include voice search, voice dialing, interactive voice response, automatic dictation of spontaneous text, telephonic access to services, voice command and control of domestic appliances. The interaction with mobile devices and other speech-based applications require the ASR system to be robust to the wide range of speakers. However, many speakers with speech difficulties are devoid of using the technology effectively

as the system models are trained with typical speakers speech. When the distorted speech is input to such systems, the system performance accuracy decreases [30,31]. The reason for low-performance accuracy maybe because the models ignore high variations (for example, nasalization, substitution errors, imprecise obstruents) in the disordered speech. A study in the literature [32,33] attempted to examine the ability of the digital speech assistants in recognizing the speech of individuals with amyotrophic lateral sclerosis (ALS)-induced dysarthria. Although the recognition accuracy is low, the researchers reported that individuals with speech impairments prefer to use the speech assistants. In this direction, another study [34] reported on how users with speech distortion prefer the use of speech-controlled devices, with many difficulties.

Additionally, in clinical interventions, studies revealed that despite the advances in surgical management and the advantages of an interdisciplinary teams involvement, a substantial number of individuals with CLP demonstrate articulation and resonance problems. For some instances, it is not easy to modify the compensatory patterns, once it becomes habituated. Intensive speech therapy is required for the improvement of such speech disorders. However, to obtain corrective-remedial interventions, individuals with CLP need access to speech therapy services. For financial, logistical, and geographical constraints the access to such services often becomes challenging. This is the reason why some CLP individuals are unable to get the effective speech therapy.

The above mentioned reason necessitates the need for improving the intelligibility of disordered CLP speech from another perspective such as signal processing techniques. The techniques such as acoustic transformations, spectral conversion, spectral tilt modification and various other spectral and temporal processing techniques are reported to enhance both the degraded speech and pathological speech. In the following section, we present the salient features of the enhancement techniques.

1.4 Overview of speech enhancement methods

Speech enhancement is a widely studied topic by many researchers in the field of speech technology because of its applications in medical, commercial, and military contexts. In the speech enhancement literature, two broad categories of enhancement studies are noted: one corresponds to the speech enhancement for degraded speech of healthy speakers and the other corresponds to the enhancement of speech sound disorders (dysarthric speech, glossectomy speech, electrolaryngeal speech, CLP, and speech enhancement for hearing impaired).

Typically, the speech signal gets degraded due to background noise, reverberation, and communication channel discrepancies. Additionally, the advancement in speech technology has increased the interest in mobile speech processing applications such as speech controlled devices, smart phone application, and assistive devices. Therefore, the methods that enhance the speech using signal processing based techniques are widely studied by many researchers. Specifically, in voice telephony, speech enhancement methods are used as preprocessing technique in speech coding or speech recognition. This section presents a review of enhancement technique.

The non-CLP speakers speech enhancement method focuses on speech in noise, where the background noise is suppressed, and the speech quality and intelligibility are improved [35]. The intelligibility of noise degraded speech is improved using the techniques, such as spectral subtraction, subspace filtering [36,37], codebook based Wiener approach [38], hidden Markov model (HMM)-based approach [39,40], dictionary-based approach [41], and deep neural network (DNN)-based approach [42]. Various other techniques attempt to improve the speech intelligibility and quality by exploiting the audio and signal properties, namely, amplitude compression scheme, dynamic range compression, and peak-to-root mean square reduction [43–45]. Certain other speech intelligibility enhancement techniques exploit the knowledge of noise mask, such as optimizations based on speech intelligibility index and glimpse proportion maximization [46, 47]. Further, adaptation based approaches are also explored to improve speech intelligibility [48]. Studies report the advantage of exploiting the naturally produced speech acoustic characteristics to improve speech intelligibility in an adverse listening environment [49, 50].

The motivation behind the enhancement of speech sound disorders is to improve the mediated interaction of pathological speaker with human or machine. In line with this motivation the dysarthric speech modification is achieved using acoustic transformation [51], Gaussian mixture modeling (GMM) based voice transformation [1], duration modification, and nonnegative matrix factorization (NMF) based spectral conversion [52, 53]. Electrolaryngeal (alaryngeal) speech enhancement includes the transformation of speech by enhancing formants and perceptual weighting technique, respectively [54, 55]. The enhancement of electrolaryngeal speech is also investigated using hybrid noise source modeling [56] and fundamental frequency control [57]. Some other studies [58, 59] have reported improvement in the quality of electrolaryngeal speech using a speaking-aid system based on voice conversion method and one-to-many eigenvoice conversion. A statistical approach is exploited in [60] to enhance

the body-conducted unvoiced speech for silence communication. Methods like frequency lowering were proposed for improving the intelligibility of degraded speech [61]. A few studies also report speech enhancement for individuals with articulation disorders using voice conversion technique [53, 62]. For hearing impaired listeners, a wide range of research is reported in the literature, where microphone arrays were shown to improve the speech intelligibility in noise. The noise reduction algorithms based on multichannel Wiener filter with and without partial noise estimates are studied in [63]. Vocoder based frequency lowering system and vowel enhancement are explored in [61, 64]. A study also focused on speech enhancement based on binary masking [65].

In the context of CLP speech where the primary source of distortion include misarticulated obstruents and vowel nasalization, the signal processing techniques are yet to be explored. Hence, motivated by the potential of speech enhancement techniques for improving SSDs, CLP speech modification is attempted in this thesis.

1.5 Motivation of the thesis

The studies mentioned in Subsection 1.3 showed that people with speech disabilities face many challenges using the speech-based technologies. However, many of them also prefer to use speech-enabled services to perform a multitude of everyday tasks with less effort. Speech-based applications provide convenience, novelty, unique solutions for the medical industry, conversational interaction, speech therapy, and learning support [66]. It is reported that some individuals with disabilities could rely on speech-based applications for additional benefits. Benefits include access to information at any time from almost anywhere, banking, weather condition, news, audiobooks, improved communication, social interaction, and navigation [67].

In the clinical intervention, even after surgery, speech distortions do not get completely eliminated. It is followed by speech therapy to achieve improvement in the speech intelligibility. Speech therapy involves speech recognition systems [68–70]. It is as effective as traditional treatment with proper monitoring by the SLPs [69]. Thus, the successful enhancement of CLP speech would be beneficial not only for the access of speech-based applications by CLP speakers but also have potential to be applied for their automated speech therapy. The outcomes of this study can be used in developing an assistive tool but not as a substitute for SLPs.

Despite the potential of the enhanced speech in the usability of speech-based applications and

speech therapy, CLP speech enhancement is not explored in the literature. Motivated by that objectives, we undertook an initial study on the effectiveness of CLP speech modification for enhancing the intelligibility and quality.

1.6 Scope of the work

The scope of the work is limited to studying specific CLP speech distortions in isolated phonemes in the context of /CVCV/ words with identical /CV/ pairs. The study primarily focuses on enhancing a few select obstruents and vowels. Later, it is extended to some nonsensical /CVCV/ words that can be formed by combining the studied obstruents and vowels. After developing the phoneme specific enhancement schemes, their impact on the intelligibility and quality of the CLP speech is studied for both phoneme-level and word-level enhancements. Further, we have also illustrated whether the proposed enhancements are effective in improving the CLP speech recognition performance.

This study has been performed on speech data collected in clinical settings from children having mild to moderate CLP disorder. Despite our attempt to facilitate CLP speakers, the scope of our study does not fully align with augmentative and alternative communication (AAC) which often focuses on severe speech distortions [67].

In the analysis of the CLP speech, we focused on the speech distortions caused by specific obstruent errors like misarticulated fricative /s/ and stop consonants /k/, /t/, /T/ and vowel nasalization /a/, /i/, /u/. In fricative misarticulation, three types of errors are addressed: shift in spectral prominence, absence of frication, and additional turbulence noise source created in the nasal cavity. All these factors lowers the required intra-oral pressure for producing fricative sounds resulting in misarticulated fricative. Fricatives represent a class of sound characterized by high-frequency energy and misarticulated fricative exhibit prominent low-frequency energy. This poses the importance of processing different frequency bands separately. In the case of misarticulated stops, absence of bursts, formant transitions, weak bursts, and weak spectral prominence are observed due to change in place and manner of articulation. Accordingly, different acoustic events around the stops must be processed separately as they possess different degrees of variation. Therefore, the acoustic events can be used as an anchor to apply different analysis and enhancement methods to improve the CLP speech. Considering the case of hypernasal speech, both the low-frequency and high-frequency spectral deviations are observed [71]. As a result the auditory perception is characterized by broadened spectral peaks

and flattened spectra [72, 73]. Each of the spectral components of the vowel has a different influence on perceived hypernasality [74]. In this study, the work is carried out on children speech, hence, the above mentioned problems gets further enhanced.

The assessment of the intelligibility of obstruent modification has been done in the context of /CVCV/ words with only target obstruent being modified. The major contributions of the thesis are:

- Development of a database for the analysis and modification of disordered CLP speech.
- Misarticulated alveolar fricative is analyzed and modified using spectral compression, emphasizing spectral tilt, and insertion method.
- Misarticulated stop consonants are automatically detected and modified using the NMF method.
- The hypernasal speech characteristics are studied, and the nasalization of the vowels is reduced using GMM based spectral conversion and temporal processing using fine weight function.
- An entire word modification is attempted based on all the isolated phoneme-specific modifications techniques.

The remaining of the thesis are organized as follows:

1.7 Organization of the thesis

- In Chapter 2 the speech intelligibility enhancement methods exploited for non-CLP speakers speech are reviewed. Taking cues from the non-CLP speech enhancement techniques, intelligibility enhancement of various SSDs are also attempted in the literature, and the same is reviewed in this chapter.
- In Chapter 3, a detail description of the data is presented. The chapter describes the data collection process, speaker characteristics, metadata, and the recording set-up. Further, it discussed the subjective evaluation performed for different speech distortions and phonemes studied in the thesis.
- In Chapter 4 the acoustic characteristics of misarticulated fricative /s/ is investigated. At first automatic segmentation is performed followed by modifying the errors close to non-CLP /s/ characteristics. Three types of misarticulated fricative /s/ are analyzed: palatalized articulation, phoneme specific nasal air emission (PSNAE) distorted /s/, and glottal stop substituted /s/. The deviations of palatalized /s/ and PSNAE distorted /s/ are corrected by modifying the spectral energy. The high-frequency energy levels are emphasized to improve the perception of

fricative /s/ using spectral tilt modification. Glottal stop substitution is modified by inserting artificially synthesized /s/. The fricative /s/ signal is synthesized using a white noise signal and linear prediction filter obtained from non-CLP children fricative /s/.

- In Chapter 5 modification of compensatory errors produced for stop consonants /k/ ,/t/, and /T/ are addressed. Three types of misarticulations are studied: glottal, palatal, and velar stop substitutions. An event-based modification approach is used to correct the misarticulated stops, where at first, automatic detection of burst onset and vowel onset events is carried out. Then, the region from vowel onset to 20 ms duration of the vowel is extracted. Further, the region from burst onset point to 20 ms duration of the vowel is defined as the region for modification, and it is transformed using the NMF based spectral conversion.
- In Chapter 6, the issues related to nasalization is studied where, hypernasal vowels /a/ ,/i/, /u/ are studied. The hypernasal (HN) speech signal spectral characteristics are modified by transforming the spectral envelope while retaining the fundamental frequency of the source speaker. The transformation is achieved using GMM based spectral conversion function derived from the source (HN) and target (non-CLP) speakers probabilistic model. Using the transformation function, hypernasal speech is mapped to the non-CLP speech. Further, a weighting function is used for de-emphasizing the interfering signal components of the XLP residual signal.
- In Chapter 7, word-level intelligibility is attempted by combining the specific phoneme modifications discussed in the above chapters. A comparison study is also done to observe whether the single transformation method exploited in the above chapters can improve the entire word-intelligibility or not. When the transformation method is used to modify the word as a whole, it is observed that the speech sounds muffled. Thus, further processing is required to achieve a good quality enhanced speech. Hence, phoneme specific modifications are exploited to improve the word-level intelligibility.
- In Chapter 8, a summary of the thesis is presented by discussing the contributions of the work. The chapter also discussed the feasible directions for future explorations based on the shortcomings of the thesis.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

2

Literature review

Contents

Overview

The correction of cleft lip and palate (CLP) speech require clinical interventions. However, the clinical strategies exploited for the structural corrections may not result in functional correction of the speech. Therefore, speech therapy is recommended. In certain cases, speech distortion persists after all the corrections are performed. Besides clinical intervention, another direction of research showed that speech enhancement can also be performed based on signal processing techniques. Several studies in the literature present speech enhancement of different speech sound disorders (SSD) for rehabilitation. Speech enhancement is also performed for non-CLP (healthy) speech, which is often corrupted by various types of environmental conditions. Therefore, taking insight from the literature studies, this work also aims at improving the CLP speech by analyzing the nature of the disordered phenomena and transforming it into non-CLP speech characteristics. The chapter first presents the review of speech enhancement techniques exploited for the non-CLP speakers degraded speech, where noisy speech, reverberant speech, and multi-speaker speech were addressed. Further, the speech enhancement performed for different SSDs such as dysarthric speech, electrolaryngeal speech, glossectomy speech, speech of the speakers after oral surgery, and speech for the hearing impaired listeners are reviewed.

2.1 Introduction

Speech tends to lose its naturalness, audibility and information content when it is corrupted by a considerable amount of acoustic background noise, interference, and recording device. The presence of noise or any form of interference causes signal degradation, and it can result in distorted speech. Speech is also distorted when it is produced by a speaker who have a speech sound disorder (SSD). SSD refers to the difficulty or a combination of difficulties with perception, motor production, and phonological representations of speech sounds [75]. The SSDs caused by an impairment affects the motor act of creating speech sounds, which may result in unintelligible speech. The degraded speech, whether it is noise distorted speech or impaired speech, can be enhanced using speech enhancement algorithms [1, 76, 77]. The purpose of speech enhancement is to improve the perceptual aspects of speech signals such as overall quality and intelligibility for listeners or machine recognition. Speech enhancement is widely studied in the literature due to a variety of applications in medical (e.g., assistive aids, speech therapy), commercial (e.g., mobile technology, navigating GPS, day-to-day service systems), and military (e.g., air force, radio relays) contexts. Accordingly, the study is also focused on one

such direction. This study aims to improve the CLP speech intelligibility and quality. The CLP speech is affected by structural and functional deformities, causing difficulty in speech sounds motor production. As CLP speech corresponds to the physiological disabilities, the study further concentrates on the speech enhancement studies performed in medical contexts. The advancements made in healthy speech enhancement studies are also reviewed to gain more insights into merits and de-merits of each techniques.

An individual with SSD may face social difficulties due to poor speech intelligibility, which can preclude them from the outside world affecting their potentials in education and employment [3]. In medical contexts, besides clinical interventions, researchers have tried to transform the disordered speech from the signal processing point of view. In a realistic environment when listeners are unable to talk face to face with speakers having speech impediments, and the speakers have to rely on certain electronic mediated application, it is assumed that a speech modification system may be beneficial [51]. The poor speech intelligibility creates problem in situations with high vocal demands, such as, telephone conversation, speaking in noisy conditions. Such situation motivates the researchers to enhance the speech for better understanding. Individuals with SSDs prefer to communicate through speech, but it is difficult to understand their speech in different situations. Even though, speech therapy has the potential to improve the intelligibility to a large extent. However, it is hard to attain the intelligibility level to that of a non-CLP speech for many of the individuals with SSDs. Therefore, it is desirable to recognize the distorted speech automatically and improve the speech characteristics. Speech enhancement systems presents one direction of research and practice, where the aim is to find ways that can improve the task that the impaired articulators should have fulfilled. Hence, to make speech communication more convenient in any situations specifically for any persons with speech pathology, it is desirable to exploit and improve the technologies that will overcome the problems.

Considering the nature of the speech disorder (i.e., deviated acoustic characteristics), a speech signal modification algorithm can be designed for enhancement. Therefore, to exploit an effective enhancement method, a detailed summary of the state-of-art signal processing-based CLP speech enhancement is required to define the work motivation appropriately. Most of the existing approaches concentrate on the intelligibility enhancement of dysarthric speech, electrolaryngeal speech, speech for hearing impaired listeners, and the speech of the speakers with oral surgery. Apart from CLP speech, enhancement of various other speech disorders has been studied to assist the needful individuals with

speaking aids and adapt better communication skills. However, works specific to CLP speech enhancement are not observed in the literature. Therefore, it is essential to study the speech enhancement performed for different SSDs. Many of the disordered speech enhancement studies benefit from the recent advances made for healthy speech enhancement techniques. In this regard, this chapter reviews the advancement made in the healthy speakers' speech enhancement methods because researchers have exploited them for improving the disordered speech quality and intelligibility.

The remaining part of this chapter is organized as follows: In Section 2.2, the advancements made in the non-CLP speakers degraded speech modification techniques are discussed. A brief discussion of speech synthesis method is presented in Section 2.3. In Section 2.4, a detailed description of different SSDs and its improvement is discussed. Section 2.6 presents the overall inferences of the enhancement and synthesis approaches exploited in different situations and for different speech disorders.

2.2 Speech enhancement techniques

Speech enhancement methods are very popular in improving healthy speakers degraded speech because of their simplicity and effectiveness. In this context, various enhancement approaches were proposed based on the type of speech distortions and applications. The techniques commonly used in speech enhancement are broadly categorized into conventional speech enhancement approaches, model-based approaches and deep learning based approaches. This section will briefly review the working principles of these approaches.

2.2.1 Conventional speech enhancement methods

The speech enhancement methods minimize the acoustically coupled background noise to improve the quality of speech communication system. In the case of speech degraded by the background noise, the noise is assumed to be uncorrelated and additive in nature. Studies report a wide variety of speech enhancement techniques because they play an important role in the context of voice telephony, hearing aids, speech coding, robust automatic speech recognition (ASR), and teleconferencing [78].

Figure 2.1 depicts the basic principle of a single channel speech enhancement technique which consider the segmented short time windowed frame as input. In the enhancement process, a short time transform is applied on the windowed frame of the observed noisy speech. Commonly used transforms include short-time Fourier transform, wavelet transform and cosines transform. The noisy speech is represented as the sum of clean and background noise given by, $y(n) = s(n) + v(n)$. Here,

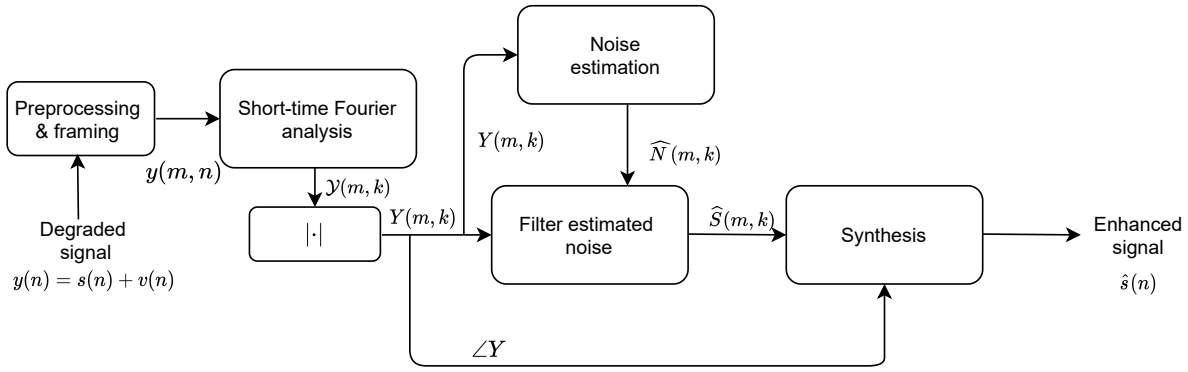

Figure 2.1: Block diagram of a conventional speech enhancement approach.

n represents the sampled time index. The variables $y(n)$, $s(n)$, and $v(n)$ denotes noisy speech signal, clean signal and noise signal, respectively. In Figure 2.1, preprocessing refers to the pre-emphasis. Further, the speech signal is segmented into overlapping frames followed by windowing. The estimate of the noise power spectrum, $\widehat{N}(m,k)$ in Figure 2.1 is acquired from the magnitude spectrum of short-time Fourier coefficient, $Y(m,k)$ of the truncated noisy speech frame $y(n,m)$. Here, m denotes the time at which analysis window is positioned and k denotes frequency index. Further, the estimated noise components are filtered to obtain the enhanced short time transform components, $\widehat{S}(m,k)$ which is then used to synthesized the enhanced signal. The phase of the noisy speech is used to synthesize the enhanced speech. Hereafter, in the context of noisy speech enhancement, phase of the noisy speech will be used during synthesis, unless specified. The above process basically illustrates the estimation and elimination of the degraded components in the speech signal.

Similar to the above described general structure of speech enhancement system, in the literature various other techniques are exploited to extract the clean speech signal from the noisy speech signal. In this regard, two subcategories were reported, namely, parametric and non-parametric approaches. In the non-parametric approach, speech enhancement is achieved by removing the noise component from the noisy observation while considering the signal distribution is unknown. The clean signal is estimated from the noisy signal under the assumption that noise is additive and uncorrelated with the clean signal. The estimate of the distortion is removed from noisy features using different classes of algorithms such as power spectral subtraction [79], Wiener and Kalman filtering [80], signal subspace filtering [81], wavelet denoising [82], binary masking methods [83], time-frequency domain approach, transform domain approaches, and schemes based on periodic models [84].

The spectral subtraction method describes the process of subtracting an average estimate of the power spectral density of the noise from the noisy speech to obtain the clean speech spectrum. While estimating the noise spectrum, some additional interfering components are observed due to errors. The interfering components change randomly and are referred as musical noise artifacts and it is perceived prominently during speech pauses. Another widely studied algorithm in speech enhancement systems include Wiener filter based approaches. In Wiener filtering based method, an optimal estimator is designed to minimize the estimation error between the clean speech estimate and the noisy speech. Wiener filtering based speech enhancement systems are reported to provide satisfactory results for stationary signals but not for the transients that occur in the frame boundaries. Another famous approach in noisy speech enhancement systems correspond to signal subspace algorithm. In this approach the noisy signal is decomposed into signal and noisy subspace respectively using singular value decomposition or Karhunen-Loeve transform [36, 37, 85]. The enhanced signal is obtained by removing the noise subspace and estimating the clean signal from the remaining signal subspace. Wavelet denoising is attempted using wavelet shrinkage algorithm. It is based on the threshold of wavelet coefficient which defines the limit between the noise and clean signal. Certain other speech enhancement systems perform denoising based on threshold. The enhancement technique separates the desired signal from the mixture by retaining the time-frequency units where, signal-to-noise ratio exceeds the pre-determined threshold while forcing to zero the remaining ones [83, 84].

In the case of parametric approach, models such as autoregressive or sinusoidal models are use to describe a signal. The clean speech is extracted from the noisy speech by formulating the noise suppression task as an estimation problem. In this type of process, the speech estimator estimates the speech spectral magnitude given the noisy speech coefficient or amplitude, variance of clean speech and noise. Both the speech and noise components are modeled as statistically independent random variables. Generally, a cost function is used to achieve the optimality criteria. In some of the studies, researchers incorporate significant perceptual information in the optimality criterion by considering log of the clean and enhanced speech amplitudes. Noise reduction using estimation process is attempted using either maximum likelihood (ML) estimator, where the aim is to suppress the noise by comparing the measured speech and noise power with the estimated background noise power [86]. The task of speech estimators such as minimum mean square error (MMSE), log MMSE, Bayesian, and short-time spectral amplitude (STSA) is to enhance speech by minimizing the error between clean and

enhanced speech [87]. The maximum *a posteriori* (MAP) estimators are used to estimate the speech spectral amplitude given the noisy speech parameters. The clean speech is obtained by maximizing the probability density function of the speech spectral amplitude [78]. Kalman filter based approaches are used for speech denoising as it can be applied to both the stationary and non-stationary signals. Kalman filter is a minimum-variance linear filter whose framework exploits human speech production system [88]. Kalman filter based speech enhancement are used in many real-world applications. The parametric approaches report improved speech quality with significant reduction in musical noise compared to non-parametric approaches. Because of the significant ability of the parametric approaches, they are found to be suitable for applications in hearing aid devices as well [89].

The conventional speech enhancement systems are widely used for noise removal and they are easy to implement. However, in these speech enhancement systems, the frequency spectra of speech and noise often overlap and noise reduction is attained at the expense of speech distortion. The accuracy of the speech enhancement system varies with the signal-to-noise ratio (SNR) and non-stationary noisy environment. Therefore, model-based approaches are developed to tackle the noise removal problems in both stationary and non-stationary noisy environments. In this type of approaches, more sophisticated statistical models are used to overcome the shortcomings.

2.2.2 Model-based speech enhancement

The model-based approaches are widely studied, where a model is considered for each of the specific types of signals (clean speech and noisy speech). In trained model-based systems, the processes are defined by parametric models (such as auto regressive models) because the parameters are estimated from the training samples of the representative databases of speech and noise. Further, based on the model parameters, a combined model is defined and the desired task (such as noise reduction, source separation) is carried out.

In the model-based speech enhancement systems shown in Figure 2.2, a short time transform is applied on the windowed frame of the observed noisy speech, which is represented as the sum of clean speech and background noise, $y(n) = s(n) + v(n)$. Here, n represents the sampled time index, $y(n)$, $s(n)$, and $v(n)$ denotes noisy speech signal, clean signal and noise signal, respectively. The power spectral density (PSD) of the speech and noise processes are estimated from the speech and noise databases, respectively. The distance between the magnitude spectrum, $Y(m,k)$ and an estimate $\widehat{Y}(m,k)$ is minimized using a distortion measure. The estimate $\widehat{Y}(m,k)$ is defined as the superposition

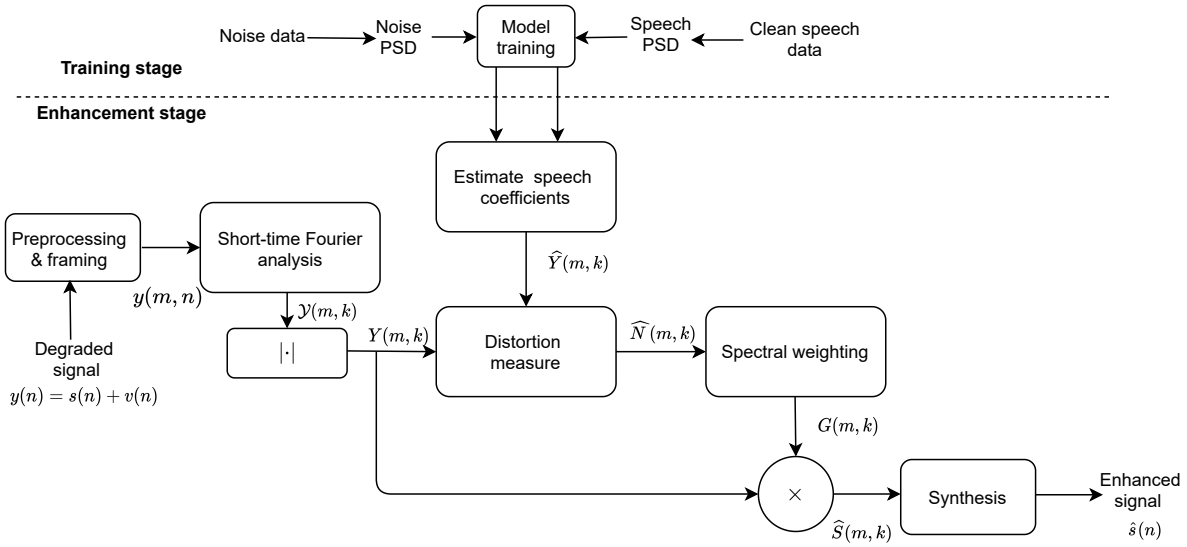

Figure 2.2: Block diagram of a model-based speech enhancement approach. PSD denote power spectral density.

of speech and noise PSD. The minimization yields a noise estimate $\widehat{N}(m,k)$. Using estimated noise PSD, spectral gain is derived. Finally, the noisy spectrum is multiplied with the spectral gain $G(m,k)$, to obtain the enhanced spectrum $\widehat{S}(m,k)$, which is further used to synthesized the enhanced speech signal.

Similar to the framework illustrated in Figure 2.2, different types of data-driven based speech enhancement techniques are developed, namely hidden Markov model (HMM) based approach, dictionary based approach and codebook based Wiener approach. Speech enhancement system designed using a HMM based approach is reported in [39]. During the enhancement process, the frame of the prepossessed noisy speech signal is analyzed for speech and non-speech activity regions. The long periods of non-speech activity regions are fed into the Viterbi-like forward algorithm and the likelihoods for each pretrained noise HMM is computed and compared with that of other HMMs. The model that exhibits highest likelihood value is considered as the representative noise model. Using the representative HMM parameters and clean speech models, the noisy speech is input to the MMSE forward algorithm. It further generates weights for Wiener filter. A single weighted filter is calculated for each frame of the noisy speech using computed filter weights and pretrained Wiener filters. Finally, the noisy speech is filtered using a weighted filter, which generates the spectral magnitude of enhanced speech signal. The modified spectral magnitude and noisy speech phase information is used to obtain the time-domain enhanced speech.

Another study reported speech enhancement using a codebook based Wiener filter [38]. The noisy speech is modified by filtering it using MMSE estimate of the clean speech. Here, at first the prior information is modeled using Bayesian MMSE estimators of speech and noise short-term predictor parameters which was developed using codebooks of linear predictive coefficients. The *priori* information is processed on a frame-by-frame basis based on the current frame observation, to ensure that the approach performed effectively in non-stationary environment. Both memory-based and memory-less estimators were developed to analyze the impact of frame-by-frame gain computation in the mean and variance of the squared error. The MMSE based approach demonstrated good performance in terms of SNR, segmental SNR, log-spectral distortion, and perceptual evaluation of speech quality. Denoising of the noisy speech signal is also attempted using immune based singular value decomposition (K-SVD) algorithm [41]. The denoising process first initializes a dictionary with a set of corrupted speech signals. Therefore, the sparse coefficients are found using an optimized algorithm based on strict sparsity constraints. The sparse matrix is assumed to be invariant while updating the dictionary atoms. Adaptive redundant dictionary is obtained by applying the dictionary and sparse representation coefficient while updating phase alternatively. Subsequently, the test set signals are decomposed over the redundant dictionary, which separates the coherent speech signals from the incoherent corrupted signal. The incoherent components are discarded to reduce the noise. Finally, the denoised speech signal is obtained by using the sparse coefficients.

With large training data and increase in computational resources, data-driven based approaches were reported to substantially advanced the state-of-the art performances. In the above mentioned approaches, speech enhancement is achieved at the expense of speech distortion. The reason may be attributed to the inaccurate utilization of the acoustic context information of the time-frequency domain. Therefore, to deal with the expense of speech distortion, researchers exploit deep learning based technologies and found that it could be more suitable to model the relationship between the clean speech and noisy observation in the time-frequency domain.

2.2.3 Deep learning based speech enhancement

Recent studies on deep neural networks (DNN) have pointed out insights on using deep learning procedure which were successfully applied to ASR and a few related tasks [90]. The tasks executed using DNN based techniques showed that it outperforms the state-of-the art techniques [29, 91]. With rapid rise in DNN and its effectiveness, researchers were inspired to investigate the ability of the

technology for speech enhancement tasks in realistic noisy environment. Speech enhancement tasks executed using deep learning approaches are considered as one of the data-driven based technology whose models are learned from the training data. The conventional speech enhancement approaches were observed to produce improved results but they accompany some speech distortion due to the inaccurate estimation of noise. Therefore, studies suggest that an adaptive and non-linear models may be able to give better performances as they will be able to model the complex relationship between the corrupted speech and noise. In this line of research, several studies employed DNN models to perform speech enhancement, such as weighted denoising encoder (WDA) [92], feedforward multilayer perceptrons (MLP) [93, 94], convolutional neural networks (CNN) [95], fully convolutional neural network (FCN) [96, 97], deep recurrent neural networks (DRNN) [98], long short-term memory networks (LSTM) [99, 100], and generative adversarial networks (GAN) [101–104].

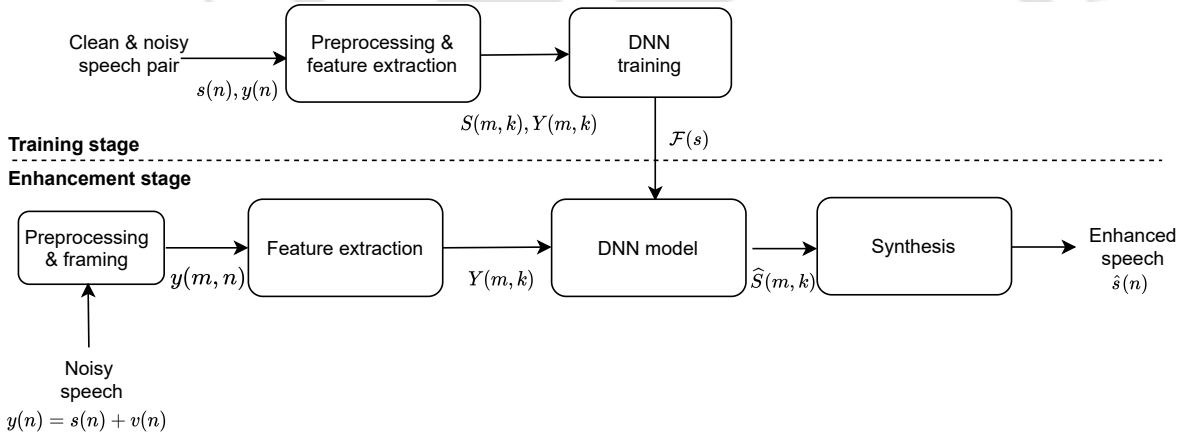

Figure 2.3: Block diagram of a DNN based speech enhancement system.

A block diagram of a DNN based speech enhancement system is depicted in Figure 2.3. In this approach, a model is trained from the log power spectral features of the pair of noisy $y(n)$ and clean $s(n)$ speech using a regression DNN model. The network is trained to develop a representative non-linear regression function $\mathcal{F}(s)$ that maps the noisy speech features $Y(m,k)$ to clean speech $\widehat{S}(m,k)$ features. The network is trained using back propagation algorithm and the parameters are updated by minimizing the prediction error using gradient based optimization method. The prediction error between the estimated and reference clean speech features can be measured using a cost function. A commonly used measure corresponds to MMSE. Further, during enhancement the noisy speech $y(n)$ is fed into the learned DNN model to generate the enhanced speech $\hat{s}(n)$. In these studies, the noisy speech phase is used while reconstructing the enhanced speech because it is assumed that phase

information is not very important for auditory perception. Rather, the magnitude or power spectra of the speech is essential and hence, more concentration is given on the estimation of the spectra to acquire the desired clean signal.

For instance in Ref. [92], WDA model was utilized to predict the power spectrum of clean speech from noisy speech. The enhancement procedure involves preprocessing (framing and windowing) the noisy speech before performing fast Fourier transform (FFT). Therefore, the normalized spectral features are input to the Gaussian mixture model (GMM) based noise classification system which is further fed to the WDA model. The WDA model is trained using stochastic back propagation algorithm where the network weights of the hidden and the output layers are adjusted using gradient-based optimizations. The WDA model is used to predict the clean speech. The prediction error between the predicted output and the clean input speech is minimized by exploiting weighted reconstruction loss function based on the spectral importance of the speech quality. Additionally, the estimated noise from noisy speech and clean speech predicted from WDA model are used to obtain *a priori* SNR. Finally, the *a priori* SNR is used to construct the Wiener filter. The clean speech spectral magnitude along with the noisy phase are used to acquire the enhanced speech. The studies in [93, 94] employed feedforward MLP for speech enhancement task. Here, speech denoising task is carried out by processing the noisy speech through the trained DNN model. An MLP is a network with feedforward connections from the input layer to the output layer, layer-by-layer, and the consecutive layers are fully connected. MLP based speech enhancement systems were observed to yield satisfactory performance in realistic noisy environments because they do not rely on prior distribution assumptions for speech and noise. They are also reported to generalize well for unknown noisy conditions.

The applicability of an MLP based network is explored for estimating ideal binary mask, ideal ratio mask, spectral magnitude mask, phase sensitive mask and complex ideal ratio mask to perform speech enhancement [105–108]. The masks in this contexts refer to the time-frequency masks that define the relationship of the clean speech to the background interference. The two-dimensional representation of the time-frequency mask is predicted from the noisy speech signal, where each of the time-frequency unit corresponds to a criteria based on the perceptual importance. The predicted time-frequency mask is multiplied with the input noisy speech to obtain the enhanced speech signal.

Although the above mentioned speech enhancement techniques showed satisfactory performance, however researchers still found room for improvements because noisy speech modification without

creating artifacts in the perceptual quality is a challenging task specially in low SNR condition and non-stationary noises. Also the above reported network architectures comprises of large number of parameters. To overcome such issues, researchers explored CNN for speech enhancement task as it is robust and represents localized low-dimensional patterns. A CNN model basically learn layers of filters to extract features and transform input data into a low resolution good representative model. Later the trained model is fed with noisy speech to generate denoised speech [95]. The neural network with deep architectures use large number of model parameters and it increases the computational load. Therefore, waveform based speech enhancement is attempted using FCN which do not require estimation of clean speech phase or using noisy speech phase. FCN works well with lesser number of parameters due to its weight sharing property. It showed better performance in terms of intelligibility and quality compared to the aforementioned enhancement methods [96, 97]. Studies reported that FCN can be used to develop a memory efficient denoising algorithm which could be implemented in an embedded device such as hearing aid [109]. Furthermore, DRNN is utilized to conduct speech enhancement due to its ability to model long-term acoustic information. DRNN architecture considers hierarchical information through multiple time scales and it makes it suitable for speech denoising task [98]. While modeling long-term acoustic contexts, limitations exists with DNN based training approaches. Hence, in this line of research, an LSTM based speech enhancement is proposed that uses recursive structure to capture the long-term contextual information [99, 100]. The combination of the above proposed techniques were also exploited for speech enhancement tasks in realistic situations.

In addition to the aforementioned speech enhancement techniques, speech synthesis methods are also used widely for speech modification in healthy and pathological speech contexts [1, 51]. Therefore, in this chapter, speech synthesis methods are also briefly reviewed.

2.3 Speech synthesis method

Speech synthesis methods approximately represent the human speech production system to a reasonable degree by simulating the speech acoustics, articulatory parameters, voiced and unvoiced excitation source characteristics. It presents a technique for generating intelligible and natural-sounding artificial speech sounds. Speech synthesis methods provide a flexible framework for speech modification as it models the vocal tract spectral characteristics (using the knowledge of speech acoustics or articulatory configurations) and excitation source characteristics independently. Hence, they can be

controlled and modified separately. Based on the type of parameters used for modeling, different forms of speech synthesis are reported in the literature, such as articulatory synthesis, formant synthesis, concatenative synthesis and statistical parametric speech synthesis.

2.3.1 Articulatory synthesis

Articulatory synthesis presents an automatic generation of high-quality synthesized speech by mimicking the human speech production process with the help of a mathematical model [110]. It models the speech production system using a complex mathematical model that involves the models of vocal tract, vocal cords, aero-acoustics, and articulatory control [111]. The articulatory control parameters include lip aperture, lip protrusion, tongue tip position, tongue tip height, tongue position and height [112]. The speech production systems are modeled with a set of area functions of uniform-tube model and the vocal cords are modeled using a two-mass model. It mainly focuses on the realistic acoustic properties of the generated speech such that particular phonemes could be represented with desired configurations of the articulators [113]. The data required for articulatory synthesis are usually derived from X-ray data or real-time magnetic resonance imaging [114].

Articulatory synthesizers along with brain-machine interface are used to constitute a tool to help the people with speech disorders [115]. Articulatory speech synthesis are also used as virtual language tutor, in speech therapy, audio-visual speech synthesis, speech encoding, text-to-speech synthesis, and speech mimicking [110].

2.3.2 Formant synthesis

A widely used synthesis method is a formant synthesis method which is based on source-filter-model of speech production system. In formant synthesis, the vocal tract transfer function is modeled by simulating the formant resonances based on specified frequency, bandwidth and amplitude [116, 117]. A set of anti-resonances are also included to model the nasal consonants and obstruents. Further, a set of rules are defined to determine the parameters necessary to synthesize the desired utterance. Such an artificial reconstruction of speech is termed as rule-based synthesis method. The input parameters include fundamental frequency (F_0), open quotient, degree of voicing in excitation, formant frequency, bandwidth and amplitudes, frequency of additional low-frequency resonator, and intensity of low and high-frequency region [118].

The formant synthesis method is incorporated in text-to-speech (TTS) synthesis systems [119]. It is

also used for esophageal speech enhancement [117], and various other speech based applications [120].

2.3.3 Concatenative synthesis

Concatenative speech synthesis system combines prerecorded individual units of speech to generate utterances that is indistinguishable from that produced by humans. Based on the type of units used for concatenation, it is divided into the following categories, namely, unit selection synthesis, domain-specific synthesis, diphone synthesis, and syllable-based synthesis [114, 121, 122]. In concatenative synthesis, the unit length affects the quality of the synthesized speech. With significant unit length, numerous units are stored in the database and accordingly, more memory is required. However, the naturalness increases with longer unit length and vice versa for shorter unit length [123].

Concatenative speech synthesis is used in TTS synthesis [123] and various other application such as speaking clock, speaking calculator, speaking weather forecast, and pathological speech enhancement [51, 124].

2.3.4 Statistical parametric speech synthesis

In statistical parametric speech synthesis (SPSS), speech is generated from the estimated speech parameters. The speech parameters include spectral and excitation parameters which are modeled using statistical generative models such as HMMs. The overall architecture consists of training and testing stages. In the training phase, acoustic parameters of the speech are extracted from the database of natural speech and model by a set of context-dependent HMMs using maximum likelihood criterion. Once the model is estimated, test data is fed into it, and based on the knowledge of the corresponding phoneme sequence, the HMMs are retrieved and concatenated to form the sentence HMM [125]. Furthermore, the speech parameter generation algorithm is used to generate the excitation and spectral parameters, which is passed through a vocoder to obtain synthesized speech [126].

All the above mentioned speech synthesis techniques provide satisfactory results and they are employed based on the desired applications. Among all the synthesis techniques, SPSS based method provides more flexibility due to its statistical modeling process. One of the main advantages of SPSS are observed in altering the voice characteristics, speaker modification (voice conversion) and emotions via transforming the model parameters [125]. The information related to speaker individuality play an important role in interpreted telephony and all the systems that make use of the pre-recorded speech namely, voice messages, dubbing, speech-to-speech translation, personalizing TTS systems,

speaking and hearing aid devices. Voice transformation usually change one or more aspects of the speech while preserving the linguistic contents and the subset of voice transformation is considered as voice conversion (VC) where one person speech is converted into that of the other.

Voice Conversion

Generally, VC refer to the study of converting one's voice like that of another while preserving the linguistic content. With the technological advancements in statistical modeling and deep learning, VC found an important role in many real-life applications. The potential applications include customizing audiobook, avatar voices, personalized speech synthesis [127], speaker de-identification [128], voice mimicry [129], disguise, dubbing, computer-assisted pronunciation training, and communication aid for speakers with impaired speech [130–132].

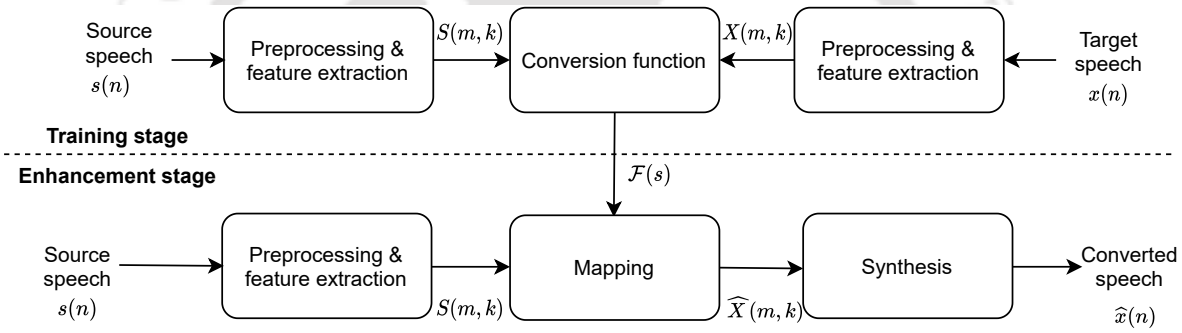

Figure 2.4: Block diagram of a generic voice conversion system.

The modules involved in the VC process is depicted in Figure 2.4. Like other data-driven based approaches, VC method also consists of training the conversion model and enhancement stages. In the training stage, the conversion function $\mathcal{F}(s)$ is trained using the features $S(m,k)$ and $X(m,k)$ extracted from the source $s(n)$ and target $x(n)$ speech. During training an objective function is optimized using the model parameters. The enhancement stage comprise of analysis, feature extraction, mapping and synthesis modules. The feature extraction procedure is similar to that followed in the training stage. Further, the converted features $\widehat{X}(m,k)$ obtained using the conversion function is used to reconstruct the transformed speech $\widehat{x}(n)$.

In the literature VC techniques are exploited in different ways based on the training data, modeling and mapping techniques. The categories are stated as follows:

- Parallel versus non-parallel training data.
- Parametric versus non-parametric statistical modeling technique.

- Frame level versus utterance level mapping.
- Monolingual versus inter-lingual conversion.

With parallel data, same utterance is considered from both the source and target speaker [133]. According to the conventional voice conversion pipeline shown in Figure 2.4, the utterances are aligned frame-by-frame using dynamic time warping method [134]. Common examples of VC using parallel training data include vector quantization [135] and GMM based VC [136, 137], partial least square regression [138], dynamic partial least square regression [139], and pitch synchronous overlap-add method [140]. Although parallel data yields significant results but in many realistic conditions, it is very difficult to obtain sufficient amount of parallel training data. Also, most of the existing VC techniques are developed for mono-lingual VC, where both the source and target speakers speak the same language. Whereas in cross-lingual VC, the source speaker and the target speaker use two different languages [141]. Parallel data is not available for such cases, hence, it renders the need of non-parallel training data to accomplish the VC task. Therefore, studies reported VC based on non-parallel training data using GANs, such as, variational autoencoding Wasserstein generative adversarial networks (VAW-GAN) [142], a variant of GAN (StarGAN) [143], and cycle-consistent adversarial network (CycleGAN) based VC [144]. Although the statistical parametric approaches were reported to perform significantly with large amount of data. However, it usually suffers from over-smoothing problem. Studies report that the over-smoothing problem can be efficiently tackled using statistical non-parametric approaches with smaller amount of training data. An effective and widely used non-parametric statistical modeling technique is non-negative matrix factorization (NMF) based VC which consider an exemplar based representation [53, 145]. The human speech production system is a highly dynamic process and the conventional frame-by-frame based VC approach constrains the modeling ability of mapping functions. Hence, adding dynamic information to the mapping features may improve the performance of VC systems [146, 147].

Recent literatures on speech enhancement studies report the use of GANs due to the several advantages relative to the aforementioned DNN based approaches. In GAN, the learning process does not require approximate inference or approximation of partition function gradient and the model has a better generalization capability. GAN is a generative model that learns to map samples from one distribution say Y into another distribution $\widehat{Y}$. In case of speech denoising task, one of the distribution Y corresponds to noisy speech signal and another distribution $\widehat{Y}$ corresponds to clean speech signal.

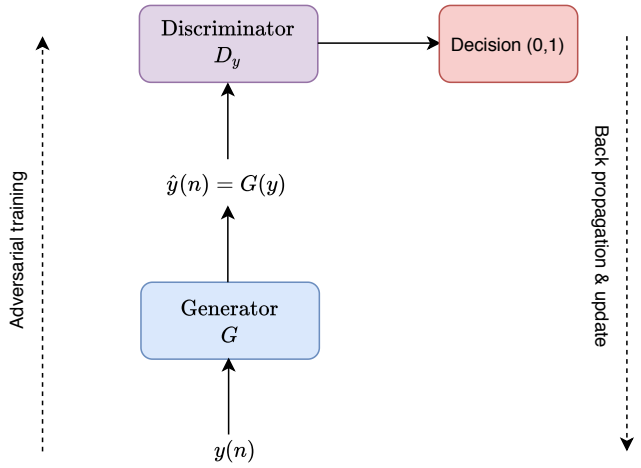

Figure 2.5: Block diagram of a generative adversarial network (GAN) model.

GAN consists of two networks : $G \rightarrow$ generator and $D \rightarrow$ discriminator and it is pictorially depicted in Figure 2.5. The network G performs the mapping that mimics the real data distribution and generates new samples related to the training data. The G network is fully convolutional, which enforces the network to focus on temporally-close correlation in the input signal and throughout the layering process. G does mapping by means of adversarial training during which it adapts the parameters of realistic data. The D network takes the output of G and provide a decision on whether the samples are real or fake. Further, the model parameters are adjusted using back propagation algorithm. Here, D gets better at finding realistic features in the input and G corrects its parameters to imitate the real data distribution. Eventually, D learns a loss function for G's output and this reinforces the G network to get rid of noisy signals that are considered as fake in this context. GAN was shown to achieve satisfactory results for source separation, singing voice separation and speech enhancement tasks [101–104]. Despite the above benefits of using GAN for speech enhancement, the need of a large amount of parallel data still lack in generalization of the whole network. Additionally, collecting a large amount of data (speech and noise) and creating a parallel corpus is challenging. This projects non-parallel VC methods more suitable for realistic situations. The CycleGAN is one of the state-of-the-art non-parallel VC methods that has shown its effectiveness for various applications [144, 148, 149].

Figure 2.6 shows the framework for speech enhancement using the CycleGAN system. A CycleGAN consists of two generators G and F and two discriminators D_X and D_Y, respectively. The generator G is a function that maps the distribution X into distribution Y, whereas the generator F

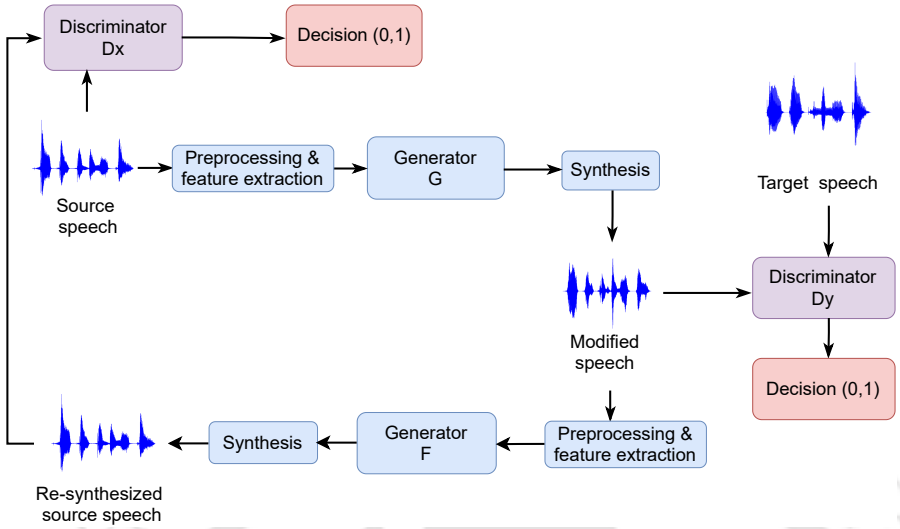

Figure 2.6: Framework for the speech enhancement using CycleGAN approach.

maps the distribution Y into distribution X. On the other hand, the discriminator D_X distinguishes X from $\hat{X} = F(Y)$. In contrast, the discriminator D_Y distinguishes Y from $\hat{Y} = G(X)$. The noisy speech serves as a source, whereas clean speech is considered as a target. Given a set of noisy and clean speech data, the CycleGAN model learns the mapping function from the training samples, which comprises of source $\{x_i\}_{i=1}^{N} \in X$ and target $\{y_i\}_{i=1}^{N} \in Y$ samples. The discriminators and the generators work collectively during training. The discriminator is trained to make the posterior probability maximum for clean/real speech and minimum for modified noisy speech. In contrast, the generator is trained to deceive the discriminator. The objective function of the CycleGAN model comprises of two losses: adversarial loss and cycle-consistency loss. An adversarial loss makes X and $\hat{X}$ or Y and $\hat{Y}$ as indistinguishable as possible. On the other hand, cycle-consistency loss guarantees that an input data retains its original characteristics after passing through the two generators. By combining both these losses (adversarial and cycle-consistency), a model can be learned from unpaired training data. The learned mappings can be further used to transform an input speech into the desired speech output. In the area of speech technology, CycleGAN is used for VC [144,150], noise robust ASR and singing voice separation [151,152]. Speech enhancement is also achieved using the combined benefits of CycleGAN and multi-objective learning [29]. The study in [29] evaluated the proposed method (combination of CycleGAN and multi-objective learning) for parallel and non-parallel data as well as unseen types of noise and shown improved performance compared to other DNN based approaches.

Inferences from healthy speech enhancement techniques for other SSDs

The speech enhancement techniques discussed in Section 2.2 are mostly exploited to improve the healthy speakers' degraded speech intelligibility and quality. The issues like denoising, dereverberation, acoustic transformations, and preprocessing the speech before transmitting it through the communication channel are addressed. The studied techniques were reported to produce better quality enhanced speech compared to the unmodified speech. Motivated by the capability and success of the existing speech enhancement methods, researchers have explored some of the reported methods for improving disordered speech intelligibility [1,51,53,153]. The techniques proposed for speech denoising are widely exploited for improving speech intelligibility for hearing impaired listeners. Furthermore, various other reported techniques are used to perform acoustic transformations, duration modification for improving dysarthric speech intelligibility and quality. Some of the speech sound disorders present with severe speech disorders, where the individuals are unable to produce certain sounds correctly, such as electrolaryngeal speech and other articulatory impairments caused by structural deficits. In such cases, the rehabilitation is attempted using synthetic speech or voice conversion techniques. In the subsequent section, speech enhancement of different SSDs are discussed.

2.4 Enhancement of speech sound disorders

For speech rehabilitation, signal processing based techniques are being widely studied to improve the speech intelligibility of various SSDs [75]. SSD is broadly classified into functional and organic SSDs [28]. Functional SSD refers to the speech impairment that impacts articulation and phonological processing. The functional SSDs have no known causes. Organic SSDs refer to the speech impairment caused by motor/neurological impairment (apraxia, dysarthria, cerebral palsy), structural impairment (CLP and other structural deficits due to trauma or surgical treatment), and sensory/perceptual impairment (hearing loss, from ear infections or other causes).

Clinical interventions are provided to the needful individuals, however, the interventions may or may not always result in correct speech articulation. Further, SLPs, evaluate the speech of the individuals who have undergone clinical interventions. Based on the type of speech characteristics, SLPs recommend speech therapy, prosthesis, communication aid, or any other surgical corrections [1, 53, 117]. Various studies in the literature reported that the modification of the disordered speech is performed based on the analysis of the disorder nature. Hence, a brief description of some of the

available disordered speech enhancement studies are discussed in the following subsections.

2.4.1 Enhancement of speech with motor/neurological impairment

The speech disorders caused by motor deficits affects the motor control of the articulators and motor programming of speech movements. Common motor speech disorder include dysarthria and apraxia of speech. Available pathological speech enhancement studies explored the dysarthric speech modification task. Hence, in this section, dysarthric speech characteristics and enhancement is discussed.

Dysarthric Speech enhancement

Dysarthria is a neurological disorder that disrupts the control of motor speech articulation. Due to the neurological damage, the speech subsystems, namely respiration, phonation, resonance, and articulation, are affected. The dysarthric speech is caused by asphyxiation of the brain, inhibiting normal development in the speech-motor areas. Different dysarthria sources include multiple sclerosis, Parkinson's disease, brain injury, stroke, Huntington's disease, myasthenia gravis, cerebral palsy, and amyotrophic multiple sclerosis [51, 154, 155]. As a result, abnormal speaking rate, muscle fatigue, and muscle weakness, intense acoustic disfluency, reduced control of articulation, reduced control of pitch, and pitch prosody are observed. All these factors lead to a decrease in speech intelligibility, making the speakers less confident while interacting with an unfamiliar listener, therefore influencing their social life.

Researchers have explored various signal processing-based techniques for the rehabilitation of dysarthric speakers, like dysarthric speech recognition systems, where the converted text is synthesized as healthy speech or voice commands. The converted text can also be used as a human-machine interaction for dysarthric speakers. Studies in the literature also reported dysarthric speech enhancement. The transformed dysarthric speech can be used for mediated human-human assistive communication, speech supportive system [156], and human-machine interaction [51].

The potential to improve dysarthric speech intelligibility is demonstrated by partial modification of the speech signal, where prosody and short-term spectra of specific speech sounds in the sentences are modified [157]. The short-term spectrum of dysarthric speech is replaced by the short-term spectra of non-dysarthric speech to obtain enhanced speech. The authors reported a 19% intelligibility improvement of the modified speech compared to the baseline system. A study in [1] reported dysarthric speech enhancement using a voice transformation system. In this work, the speech transformation sys-

tem is studied in consonant-vowel-consonant (CVC) contexts. The enhanced speech signal is obtained by transforming the short-term vowel spectra by mapping the features towards the non-dysarthric target features utilizing the learned transformation function. The transformation function is learned using a GMM. Before modification, the formants are annotated manually. The final output signal is obtained by concatenating the transformed voiced segments with original unvoiced segments. A vowel identification task is used to evaluate the modified signals, and it showed a significant improvement in speech intelligibility.

Another study on dysarthric speech enhancement presented a serialized sequence of acoustic transformations. Each of the transformations is designed in response to the unique effect of the dysarthria on speech intelligibility [51]. The speech modifications in this study are carried out using the TORGO database. The acoustic transformations include high-pass filtering the unvoiced consonants, correction of insertion and deletion errors, tempo morphing, and frequency morphing. The transformed speech signals evaluated experimentally using human listeners, and an automatic speech recognition system showed significant improvement of the modified dysarthric speech compared to original unmodified speech.

Several other researchers report the dysarthric speech intelligibility enhancement using the TORGO database, where one of the studies addressed the devoicing error modification [158]. The authors carried out the study by first segregating the region for modification by automatically detecting the acoustic landmarks, followed by inserting a voice bar in the stop-closure region. The dysarthric speech quality improvement based on durational analysis is reported in [52]. The performance is evaluated using the Nemours database and speech data collected from a dysarthric speaker of Indian origin. The authors report significant improvement in speech quality after modification.

In a different study, the dysarthric speech intelligibility improvement is attempted for ten dysarthric speakers in the Nemours database [156]. The study first analyzed and identified each of the dysarthric speakers articulation errors, using an isolated-style phonemic recognition system trained with TIMIT speech corpus. It is followed by a likelihood Gaussian-based analysis. Based on the speaker-specific dictionary and bigram language model, the estimated articulatory errors are incorporated into a phoneme recognition system. Finally, the error-corrected text is synthesized using a HMM-based speaker-adaptive speech synthesis system. The speech rate of synthesized speech is further subjected to speech rate modification using the time-domain pitch synchronous overlap-add method. The authors

report that it further enhances the naturalness of moderate and severe dysarthric speech.

In [159] and [160], authors reported consonant enhancement for the speech that is degraded due to cerebral palsy. The motor disorders of cerebral palsy affect the movement of articulators, resulting in motor speech disorder dysarthria. The speech enhancement is achieved using an exemplar-based spectral conversion using a NMF method. Here, the source speaker spectrum is converted into a target speaker spectrum. The evaluation results indicate that the NMF method improved the quality and clarity of the consonants significantly.

2.4.2 Enhancement of speech with structural impairment

Structural impairment affects speech production due to inadequate build-up of requisite air pressure in the oral cavity or inability to produce speech sounds due to removing a part of the articulatory system. Speech disorders namely, electrolaryngeal speech, glossectomy speech, and speech disorders resulting after oral surgery caused by structural deficits, are addressed in the literature. A brief discussion of the same is presented below:

Enhancement of electrolaryngeal speech

Individuals who have had the entire larynx removed after laryngeal cancer experience loss of ability to produce speech as desired. An early symptom of laryngectomy is observed as the degradation in the voice quality [161]. Due to the removal of the larynx, phonation and other speech subsystems are affected, resulting in the inability to produce natural voice.

Several types of research indicate that most of the laryngectomees rely on electrolarynx (EL) as their primary communication method for voice rehabilitation. Due to the ease in learning, operation and continuous output, EL is mostly adopted by the individuals with laryngectomees. An EL is a hand-held battery-powered device, which is pressed against the neck to transmit the electronic sound into the oral cavity. The electronic sound is generated by using an electromechanical vibrator. The electronic sound source is transmitted through the tissues of the neck into the oral cavity, and the user modulates the sound source via the movement of the articulators to generate speech. Although EL speech is preferred by most of the laryngectomees and it has been clinically proven to be an essential method of vocal rehabilitation, however, the poor intelligibility and quality of the EL speech limit the application of EL. The use of EL exhibits many problems. Therefore, researches were mainly focussed on the characterization of phonation disorders in order to design techniques for its correction.

Signal processing based techniques are explored for the intelligibility and quality improvement

of EL speech. In [54], alaryngeal/electrolaryngeal speech enhancement is attempted using a speech conversion algorithm (vector-quantization and linear multivariate regression), where the algorithms are modified to reduce the spectral distortion and spectral discontinuity. The spectral distortion was reduced by enhancing formants using chirp Z-transform, and spectral discontinuity was corrected using overlapping clusters during the conversion mapping function training. The results of the study revealed that the listeners preferred the modified alaryngeal speech over the original speech. Authors in [55] reported use of a perceptual weighting technique to adapt the subtraction parameters, which effectively reduces the radiated noise of electrolaryngeal speech. The subtraction parameters consist of the subtraction factor and spectral floor. The subtraction parameters are used to reduce the noise parameters from the noisy speech to obtain enhanced speech. Some other studies in [58, 59] improved the quality of electrolaryngeal speech using a speaking-aid system based on voice conversion method and one-to-many eigenvoice conversion.

Similarly, the statistical approaches are exploited in [60] to enhance the body-conducted unvoiced speech for silence communication. In a study, the monotonic EL speech is addressed using a fundamental frequency (F0) control method [57]. The study was carried out using monosyllables, disyllabic words, and frequently used phrases in Mandarin EL speech. The authors presented a touch-controlled electrolarynx prototype. The study reported that the touch-controlled electrolarynx output closely matches the healthy speech pitch contours corresponding to the four Mandarin tones. In one of the studies, researchers minimize fricative distortion caused by improper EL source and deviant physiological structure of the vocal tracts [56]. To improve the fricative characteristics, a hybrid noise source is proposed in the study, which is obtained by combining the healthy speaker's fricative sources and the laryngectomee speakers compensation source. The compensation source for the fricative of a laryngectomee speaker is referred to as the acoustic defects observed in the frequency domain, which occurred due to improper source transmitted through the neck generated by the EL. Five Mandarin fricatives are considered in the study. Authors reported that hybrid noise source significantly improves the intelligibility of EL fricatives by improving the spectral shapes and altering the spectral energy concentration region.

Enhancement of speech distorted after oral surgery

Speech distortion is often observed after an individual has been through an oral surgery, where either a tongue or a part of the articulator is removed during the surgical treatment of the vocal tract system.

The removal of part or whole of the tongue is referred to as glossectomy, and it affects the speech sound productions severely [162]. The individuals with glossectomy face considerable difficulties producing vowels, consonants, and fricatives, respectively [163, 164]. Similarly, many surgical patients have had parts of their articulators removed due to several other oral problems [53, 165]. The speech production system is severely impacted. As a result, such an individual's speech is often difficult to understand by an unfamiliar listener.

In the literature, several approaches were reported to improve the intelligibility of the speech of patients who have had parts of their articulators removed during surgery. After oral surgery, most of the individuals have a structural deficit in the articulatory system, which hinders them in producing certain speech sounds. Therefore, researchers tried to improve the speech intelligibility by using voice conversion techniques rather than correcting the disordered speech signal characteristics.

In [53], a joint dictionary learning-based non-negative matrix factorization algorithm is exploited to improve the disordered speech. Here, the algorithm simultaneously learns the basis of source and target spectra. While learning the bases spectra, a small number of bases is specified. The algorithm learns a set of bases that represent the entire set of examples. The motive of specifying the number of bases is to improve the conversion efficiency. Using short-time objective intelligibility scores, the authors demonstrated that their proposed method achieved higher scores than the original unconverted speech. In [162], intelligibility improvement of a wide glossectomy and/or mandibulectomy speech is attempted. Mandibulectomy is the surgical procedure where all or part of the jaw is removed. Here, a GMM based voice conversion method is used to correct the distorted speech signal. The authors demonstrated that the converted speech mel-cepstral distance is decreased by 40% compared to unconverted speech. The authors also demonstrated that using GMM-based voice conversion method, they are able to reconstruct the high-frequency spectra of the phonemes /h/, /t/, /k/, /ts/, and /ch/ successfully. A study is performed in a similar direction to improve the naturalness of the reconstructed glossectomy speech, being generated using the voice conversion method [62]. The researchers exploited the spectrum differential method to modify the waveforms. In the study, the authors showed that the power in the high-frequency region of the reconstructed fricatives and stops exhibit similar characteristics to that of the healthy speakers speech. The improvement of articulation disordered speech is addressed using an end-to-end GAN based unsupervised VC model [166]. The approach transforms the disordered speech into that of healthy speech while retaining the linguistic information

and speaker characteristics.

2.4.3 Enhancement of speech for sensory/perceptual impairment

The perceptual impairment refers to the complete or partial hearing loss. The aspects of severity may vary from person to person, however, mild hearing loss may cause difficulty in understanding the speech in different situations, especially in noisy conditions. Individuals with moderate to severe hearing loss may need a hearing aid. People who rely on hearing aids try to acquire healthy speech characteristics and perceptivity. However, in certain demanding conditions, they still face difficulty in understanding the messages. This further provides a room for researchers to utilize signal processing algorithms and adapt the speech perceptivity according to the desired situations. Hence, this section discusses the works being carried out for enhancing the speech for hearing impaired listeners.

Speech enhancement for hearing impaired listeners

Hearing impairment is considered as one of the most prevalent communicative disorder. Hearing impairment or deafness is described as the inability to perceive speech sounds. As a result, an individual with severe hearing loss often finds difficulty learning all aspects of a language [2]. Due to preliminary hearing difficulties, hard of hearing listeners face a significant problem in understanding speech in everyday communication in a noisy environment.

Hearing impaired listeners often require a higher signal-to-noise ratio than normal-hearing listeners to perceive the spoken message correctly. Accordingly, several signal-processing strategies: frequency lowering, noise reduction algorithms, vowel coding in the auditory system, array processing techniques were developed to improve the perceptivity for hearing impaired listeners. Some studies report that suppressing interference from other sources rather than the desired signal direction leads to improvement in speech intelligibility.

A study in [167] demonstrated that microphone arrays could improve the speech intelligibility in noise for hearing impaired listeners. In this study, the researchers evaluated the speech intelligibility using two array processing techniques, delay-and-sum beamforming and super directive processing. The speech intelligibility measured using speech reception threshold and speech intelligibility rating showed that super directive processing results in significant intelligibility improvement. In [63], researchers focussed on the evaluation of speech enhancement in binaural multimicrophone hearing aids. Authors in the study considered noise reduction algorithms based on multichannel Wiener filter and multichannel Wiener filter with partial noise estimate. They evaluate the noise reduction algorithms in

different speech-in-multi-talker-babble noise scenarios. The authors concluded that the binaural multichannel Wiener filter-based algorithm offers an alternative standard adaptive directional microphone in a realistic acoustic environment.

In [61], a vocoder based frequency lowering system, is described where the spectral differences of fricatives are enhanced. Perceptual evaluation by normal-hearing listeners showed that the proposed system results in improved perceptivity of fricatives and affricates. A study developed an algorithm to separate speech from noise based on binary masking [65]. Unlike the ideal binary mask, the authors have estimated the mask by training the algorithm over the data that were not used during testing. The intelligibility evaluation is performed using normal and hard of hearing listeners, and they indicate intelligibility improvement of the processed speech signals. To improve the speech intelligibility for hearing impaired listeners, vowel enhancement is explored in [64]. The study demonstrated a strategy to restore vowel encoding at the level of the auditory midbrain. The signal processing based approach consists of pitch tracking, formant tracking, and formant enhancement. The subjective listening test infers that the modified sounds are different and acceptable compared to the original sounds. However, additional noise artifacts are also observed in the modified speech. Hence, the authors stated that the system is not suitable for real-time use, but it can be exploited for testing. Frequency lowering technique is also exploited in another study [168], where the nonnegative matrix factorization method is used to improve Mandarin speech recognition. Authors have affirmed that the proposed system can be utilized across different languages. The authors conclude that the proposed method performed significantly better for affricates and fricatives compared to stop consonants.

A work in [169] reported the use of a smartphone as an assistive device for hearing impaired listeners. The study developed a single microphone speech enhancement technique based on the super Gaussian joint maximum *a posteriori* method. The speech enhancement technique incorporated a controlling parameter that allows the user to adjust the noise suppression and speech distortion amount based on their hearing comfort. The experimental evaluation results supported the effectiveness of the work. In [170], a multi-objective learning based deep denoising autoencoder, is proposed to improve the perceptivity of hearing impaired listeners in noisy conditions. The evaluation results showed that the proposed method effectively reduces the background noise compared to the deep denoising autoencoder approach. In [171], the improvement in intelligibility and naturalness of the speech is performed for deaf speakers by adapting a pre-trained normalization model. Here, an end-to-end-

trained speech-to-speech conversion model-Parrotron is used to map the source spectrogram into a target spectrogram. The authors demonstrated that the model can be adapted to normalize the speech of a deaf speaker. The converted speech showed significant improvements in intelligibility and naturalness that were measured using ASR and perceptual listening tests.

2.5 Scope of existing techniques for CLP speech enhancement

Based on the discussed speech tecnhiques above, the salient features of the methods are discussed while highlighting their feasibility for CLP speech enhancement.

- **Normal speech enhancement approaches:** The speech enhancement approaches for healthy speakers speech mentioned in Section 2.2 were proposed for noisy speech modification, assuming that noise is additive and uncorrelated with a clean signal. In order to use the speech enhancement methods to improve CLP speech, which were proposed for noisy speech modification, it is required to acquire the following aspects:
  - Nasal spectrum from the high pitched speech and remove it from the hypernasal spectrum to obtain a non-nasal spectrum.
  - Nasal and non-nasal amplitude.
  - Decompose the vector space of hypernasal speech into non-nasal and nasal subspace.
  - Other nasal and non-nasal speech parameters.
  - Decompose the misarticulated and non-misarticulated speech components.

  However, unlike noisy speech, in hypernasal speech, nasality is not a distinct component from speech. It is produced from the same vocal tract system, and it is correlated with speech. Similarly, it is also not possible to decompose the misarticulated sounds and non-misarticulated speech components as the misarticulations are produced from the same vocal tract system. Hence, for the modification of speech sound disorders, it is essential to understand the nature of the pathology and then use the enhancement method accordingly.
- **Speech synthesis approaches:** Considering the speech synthesis approaches discussed in Section 2.3, the following inferences are derived.
  - **Articulatory synthesis:** With articulatory synthesis, optimizing the mathematical models and characterizing the 3-dimensional nature of the vocal tract will be very complex, and challenging for the speakers with CLP. Additionally, specialized apparatus is required

for collecting articulation data, which is not practical for real-time applications. Moreover, when the speakers with CLP already have articulatory issues, then the output synthesized speech will result in a disordered synthetic speech, which will not serve the purpose of the study. Hence, articulatory synthesis is not suitable for this study.

- **Formant synthesis:** For the speakers with CLP, due to the impairment in the craniofacial region, it is challenging to design rules specifying the source and vocal tract parameters, which will result in de-nasalized speech or non-misarticulated speech. Also formant synthesis method does not model any physical characteristics of vocal tract system.
- **Concatenative synthesis:** It is reported to produce natural-sounding speech, but the speech quality varies with the unit length, under-represented phonemes in the database, bad joints, prosody, and size of the speech corpus. With the help of unit selection synthesis, the same speaker voice can be used to generate the speech. Since the CLP speech is already distorted, some transformation need to be applied to mimic the synthesized speech close to natural speech. Studies showed that the output of a concatenative synthesizer is used to create training data for the voice conversion model [172]. Therefore, unit selection synthesis system may not be suitable for this work. TTS synthesis system designed by the concatenative synthesis method generates highly natural speech, but it is difficult and costly to synthesize high-quality speech with various voice quality [126]. The quality of the output speech depends on the pre-recorded speech with limited voice, size of the speech corpus, which requires large memory. Therefore, it is not feasible to modify the speech characteristics and preserve some individuality information of each of the CLP speakers.
- **Statistical parametric speech synthesis:** The quality of synthesized speech is quite different from natural speech as it exhibits artificial sound and muffles. Although speaker adaptation is possible with SPSS, but due to the limited number of speakers that a synthesis system can use, personalized TTS is difficult and expensive to obtain [126]. The naturalness of an SPSS based systems exhibit deficiencies in accurately mimicking the naturalness of that of the target speaker. In the case of DNN-based statistical synthesis, it efficiently estimates the acoustic models with complex context dependencies, and the speech quality is very good with a sufficient amount of data.

- **Enhancement of SSDs:** In Section 2.4.1, dysarthric speech improvement is discussed. Specific

enhancement techniques are employed based on the disordered phenonmena, such as vowel distortion, devoicing error, deletion error, insertion error, and an inadequate speaking rate. For correcting such errors, techniques based on voice transformation, synthetic speech, insertion of voice bar, and nulling the spectral energy in specific frequency bands are employed. From the analysis of deviated acoustic characteristics of CLP speech, it is observed that some of the distorted speech characteristics are similar to that of dysarthric speech. Hence, some insights can be drawn from the dysarthric speech enhancement methods for CLP speech enhancement. For example, in the devoicing error analysis of CLP speech, it is observed that in addition to the low-frequency voiced component in the closure region, the burst and the voice onset time of the devoiced stop are distorted in some cases. Unlike in dysarthric speech enhancement, only the insertion of voice bars may not result in CLP speech correction. Therefore, along with voice bar insertion, burst, and voice onset time modification must be attempted to obtain enhanced speech. Furthermore, voice transformation is exploited to modify the distorted vowel formants in dysarthric speech. Similarly, vowel distortions due to nasalization are observed to impact CLP speech perceptivity significantly. Hence, de-nasalization of the vowels in CLP speech may be achieved using the voice transformation method reported for dysarthric speech. Deletion error is addressed in the dysarthric speech. A similar kind of error is also observed in CLP speech. The absence of phoneme due to glottal stop error and other articulation errors are observed. Hence, in such a case, the insertion of synthetic speech may be exploited for its correction as it was performed for dysarthric speech correction. In a dysarthric speech, before modifying the articulation errors, they are automatically identified using a recognition system. The error-corrected text is generated using a synthesis system. In a similar direction, automatic identification of CLP speech articulation errors followed by its correction can be explored for CLP speech modification. Additionally, to overcome the smoothening effect caused by GMM based voice conversion, NMF based voice conversion was exploited to improve the intelligibility of distorted consonants in dysarthric speech. Consonant misarticulations are frequently observed in CLP speech. If GMM based voice conversion is used for its correction, the smoothing effect may hinder the consonants improvement. Hence, NMF based voice conversion can be explored for its correction.

From the discussion of electrolaryngeal speech improvement in Subsection 2.4.2, it is noted that

various modification techniques were employed to reduce the radiated noise generated from the use of electrolarynx. Hybrid noise source modeling and controlling fundamental frequency are also used to obtained enhanced speech. The reported methodologies are useful in a situation when noise is assumed to be uncorrelated from speech. Hence, the estimated noise can be subtracted from the noisy electrolaryngeal speech. However, for CLP speech enhancement, such noise reduction techniques may not result in good quality enhanced speech because the distortions in CLP speech are correlated with speech produced from the same vocal tract system. If CLP speech enhancement is realized in a real-time environment, the noise reduction technique may be useful. In some cases, statistical approaches like voice conversion methods are exploited to transform the distorted electrolaryngeal speech. In such cases, a model is trained using the input from both the source (electrolaryngeal) and target (healthy) speakers speech, and the trained model is further used for speech conversion. The voice conversion technique can be employed for CLP speech correction because the distorted CLP speech can be modified by mapping it onto the healthy speech.

For the rehabilitation of the individuals whose speech is distorted after oral surgery, researchers have employed the voice conversion method mainly because removing a part of the articulator precludes them from producing certain speech sounds. Therefore, mapping the acoustic characteristics of the speech using voice conversion techniques is an efficient way of speech modification for individuals with oral surgery. In CLP speech, the speech is often degraded due to articulation errors like glottal stop substitutions, nasal substitutions, and several other pharyngeal substitutions. To modify such errors in CLP speech, voice conversion techniques may result in enhanced speech.

From the speech enhancement studies performed for hearing impaired listeners, insights were drawn for CLP speech enhancement. The noise suppression techniques can be exploited for CLP speech enhancement if the system is used in real-time applications. Except for the noise suppression techniques, other modification techniques like frequency lowering techniques and dynamic range compression, spectral energy shifting may be useful for CLP speech enhancement because, in CLP speech, errors like palatalization of fricatives are observed. In such cases, the spectral energy concentration needs to be shifted from the lower frequency region to the higher frequency region to correct the misarticulated fricative. Hence, such approaches can be explored

for CLP speech enhancement.

Based on the type of speech disorders demonstrated by the CLP speakers i.e., nasalization, substitution error and glottal stops, "statistical parametric speech synthesis" forms the most suited method for CLP speech enhancement. Additionally, various signal modifications are also applied to generate intelligible CLP speech. Therefore, some of the speech enhancement algorithms used in the thesis fall into "enhancement of speech sound disorders" category.

2.6 Summary

In this chapter, the reported speech enhancement studies performed for healthy speech and disordered speech are discussed. The merits and demerits of the explored approaches and the scope for CLP speech enhancement are also presented. It is noted that depending upon the nature of the speech disorder spectral transformation techniques and signal processing based methods can be exploited. The disordered speech enhancement studies presented in the literature do not make any attempt for CLP speech. Furthermore, the dysarthric speech, glossectomy speech, electrolaryngeal speech, speech enhancement for hearing impaired listeners, and other disordered speech are studied in the literature. The commonly studied disordered speech exhibit different speech characteristics compared to CLP speech, where palatalization, nasal air emission, glottal stop, nasalization of voiced stops are observed in CLP speech. The existing methods for disordered speech enhancement performed effectively for the reported errors. However, they may or may not be useful for CLP speech modification.

Taking insights from the literature of speech enhancement methods for healthy speech and disordered speech, in this thesis, different attempts were made to improve CLP speech intelligibility. In the clinical environment, correcting functional articulation errors are considered one of the primary goals of treatment because increasing articulation capability of phonemes enhances speech intelligibility. Each of the speech disorders has a different impact on speech intelligibility. Hence, the SLPs emphasize on the phoneme based assessment and articulation therapy for CLP speech enhancement. Therefore, the production-based knowledge is studied in the study, and accordingly, modifications are performed.

The thesis aims to perform CLP speech intelligibility enhancement by comparing it with the acoustic characteristics of the healthy speech. The deviated acoustic characteristics of the CLP

speech will be analyzed and subject to modification. It is speculated that modifying the deviant acoustic characteristics of CLP speech will lead to good quality enhanced speech because the articulatory impairment causing speech distortion will be modified.

3

CLP Speech Database Development

Contents

Overview

This chapter discusses the development of cleft lip and palate (CLP) speech database. The database includes age and gender matched CLP as well as non-CLP (healthy) children speech. The non-CLP children data are considered as control subjects. All the speakers considered in the study are native Kannada speakers. The chapter provides a detailed description of the database including the speaker details, speech stimuli, assessment procedure, and phonemic annotations. This chapter also includes detailed description of the type of disorders, speakers, and speech samples exploited in the subsequent chapters. Further, inter-rater reliability study is performed for three different categories of explorations being carried out in the thesis.

3.1 Introduction

The speech distortions demonstrated by the individuals with CLP reduces the speech intelligibility. To improve the intelligibility of CLP speech, a detailed assessment is performed by the speech language pathologists (SLPs). Based on the assessment, the SLPs recommend for any kind of clinical interventions, be it surgery, prosthesis or any type of behavioural therapy. Prior works in the literature have indicated that besides clinical strategies, another direction of rehabilitation for pathological speakers can be acquired by performing speech enhancement using signal processing techniques. Thus, motivated by the literature works, this study performs CLP speech enhancement. To carry out the study, CLP speech database is required and it is not freely available in the public domain. Hence, this chapter discusses the acquisition of a new database of Kannada speaking CLP children.

The database was created in collaboration with the SLPs of All India Institute of Speech and Hearing (AIISH), Mysuru, India. The database consists of age and gender matched CLP and non-CLP speakers' speech. The non-CLP speakers served as controls for the study. Both the CLP and non-CLP children are native Kannada speakers. Prior to the recording, ethical consents were obtained from the parents/caregivers of each of the participants. An overview of the study is provided to the parents/caregivers. The study was conducted with clearance from the AIISH Bio-behavioral ethical committee.

The rest of the chapter is organized as follows: Section 3.2 provide a description of the data collection procedure where the speaker details, design of the speech stimuli, and labeling of the speech

data is described. In Section 3.3, the speech assessment process for all the speech samples that are included in the study are described. It is followed by the description of non-CLP speech data in Section 3.4. In Section 3.5, a brief summary of the chapter is presented.

3.2 Data collection

Data collection from the speakers with CLP is quite challenging because they are susceptible to fatigue and have varying degree of speech intelligibility. As a result, it is difficult to acquire enough example of the utterances from them. Therefore, for this study, CLP speech data is collected by considering various aspects of the speakers and the corresponding speech disorders.

3.2.1 Speaker details

Each of the CLP participant had either a repaired CLP or a repaired cleft palate. The children with CLP had undergone one or more surgeries, namely, repair of cleft lip, repair of cleft palate, pharyngeal flap surgery, sphincter pharyngoplasty, tonsillectomy and/or adenoidectomy, secondary surgeries to repair palatal fistulae, alveolar bone grafting, and surgical treatment of malocclusion. The term repaired cleft palate is used to refer to the individuals who had undergone surgical intervention for the palate. The recorded data comprised of speech data with mild, moderate and severe speech distortions. Accordingly, the database consists of varying degrees of intelligibility. The severity grading of the speech disorder is done by the expert SLPs. None of the individuals with CLP had any history of other developmental difficulties. Each of the individual with CLP had adequate language abilities. The CLP participants were recruited by the SLPs at the AIISH. Forty-two CLP speakers (18 females, 24 males) in the age range of $7 - 12$ years have participated in the study.

Apart from the CLP speakers, the data from forty-two non-CLP speakers is also collected in the same recording setup. The description of the same is provided later in this chapter. The non-CLP speech data is used as a reference for the correction of the distorted CLP speech. In the case of spectral conversion, the non-CLP speech data is used as a target to which the CLP speech is mapped to.

3.2.2 Recording setup

The SLPs of AIISH have designed suitable meaningful words, non-sensical consonant-vowel-consonant-vowel (CVCV) and vowel-consonant-vowel (VCV) words, sustained phonations of vowels, phrases, and sentences to evaluate the speech outcome measures of the individuals with CLP. While designing the

speech stimuli for recording, the SLPs have tried to cover a range of acoustic contrasts with phonetically balanced phrases and sentences. In this thesis, non-sensical /CVCV/ words and vowel phonations are used for analysis and modification. All the speech samples are recorded in a sound-proof room using a speech level meter (Bruel and Kjær) at a sampling frequency of 48 kHz and 16-bit resolution [173]. The recorded speech samples are saved in a .WAV format. The microphone was placed at a distance of 15 cm from each speaker while recording. During recording, the instructor first uttered the target word, then the response of the children was recorded. Each of the expert SLP had an experience of around five years in the field of CLP speech evaluation. The speech samples are recorded for $2-3$ sessions for each of the speaker.

3.2.3 Metadata

After recording, all the speech samples are manually segmented into a sequence of phoneme labels with time alignments. The database consists of the following type of metadata:

Nonsensical words: These words are used to gauge the articulatory precision of obstruents in the presence of low, mid and high vowels. In non-sensical words, /CVCV/ and /VCV/ words are considered. The /CVCV/ word spans the space of all possible speech sounds used in CLP speech assessment by forming a combination using one of the phonemes, /k, g, ch, j, T, D, t, d, p, b, s, sh, m, n, N, r, w, l, y/ in the vowel contexts /a/, /i/, and /u/, respectively. In case of /VCV/ word, the V unit corresponds to vowel /a/ and it is combined with one of the aforementioned phonemes. Through the non-sensical words, the SLPs basically intend to observe the phonetic contrasts between the obstruents and vowels. The /CVCV/ and /VCV/ words are helpful in studying the formant transitions between vowels and consonants, spectro-temporal characteristics of stop consonants, and the concentration of spectral energy in certain frequency bands. The phonetic contrasts may get influenced due to nasalization and other articulatory impairments.

Table 3.1: Description of nonsensical words.

kaka	gaga	chacha	jaja	TaTa	DaDa	tata	dada	lala	rara
kiki	gigi	chichi	jiji	TiTi	DiDi	titi	didi	lili	riri
kuku	gugu	chuchu	juju	TuTu	DuDu	tutu	dudu	lulu	ruru
papa	baba	sasa	shasha	mama	nana	NaNa	wawa	yaya	
pipi	bibi	sisi	shishi	mimi	nini	NiNi	wiwi	yiyi	
pupu	bubu	susu	shushu	mumu	nunu	NuNu	wuwu	yuyu	

Meaningful words and phrases: The meaningful words and phrases are used to study the acoustic characteristics of the consonants and vowels in natural and uncontrolled contexts. The meaningful words are useful in evaluating the type of error that are influenced by the phonetic context.

Table 3.2: Description of meaningful words.

paTa	kasa	papu	kiTaki
baDa	gata	tabala	bassu
Tapa	saja	koti	sara
Daka	chaTa	gooDu	kaaDu
taTa	jaDa	paTaki	chape
daDa	kaidi	daBBi	saude

Table 3.3: Description of phrases rich in oral consonants.

O1	kage kalu kappu	O7	sarita kattari taa	O13	chacha chaati kodu
O2	geeta bega hogu	O8	idu hosa batte	O14	adu chopada chaaku
O3	dana daari tappitu	O9	shivana uru kaashi	O15	ajja ajji jaatrege hodaru
O4	appa paTa ta	O10	sharada shalege hodalu	O16	adu joga jalapata
O5	baalu tabala barisu	O11	sumaa sara koDu	O17	paTa paTa bavuta
O6	beDa kaaDige Odida	O12	bisi bisi gasa gasa payasa	O18	adu daapa davasa
O19	taata tabala taa				

Table 3.4: Description of phrases rich in nasal consonants.

N1	manu aneyannu nodida
N2	naveena maneyinda bandanu
N3	nanu aneyannu noDide
N4	manga maneya melide
N5	mama mandyadinda bandide
N6	maamana mane mangaloorinallide
N7	meenalige negadi bandide
N8	nari neladinda negeitu

These types of speech stimuli are designed in such a way that they are loaded with pressure

consonants targets and they do not include any nasal consonants. The meaningful words and phrases are used to assess hypernasality, nasal air emission, and consonant production errors. In clinical settings, the meaningful words and phrases are important to analyze because an individual with CLP sometimes produce a phoneme correctly in isolation, but produce it as different type of error at word-level and sentence-level due to the influence of the phonetic context in casual conversation. The recorded nonsensical words, oral and nasal phrases are shown in Tables 3.2, 3.3, and 3.4 respectively.

Sustained vowel phonations: In the individuals with CLP, velopharyngeal dysfunction results in air-leakage through the nasal cavity during phonation leading into hypernasality. Studies report that as hypernasality is caused by abnormal nasal resonance of sounds, it is associated with the phonated sounds [2]. Therefore, the acoustic deviations caused by hypernasality are studied using sustained vowel phonations in consonant-vowel (CV) structures. The sustained vowel phonations are relatively free from the influence of phonetic contexts, intonation, stress, and speaking rate. The database consists of three vowel phonations: /a/, /i/ and /u/.

3.2.4 Data labeling

This work aims at analyzing and improving the obstruents, vowels and the transition regions. Hence, for this purpose, the burst, vowel onset points, onset and offset of fricatives are manually labeled via a careful visualization of the speech waveform and spectrogram using PRAAT software [174]. The phonemic annotations are carried out by a person having the knowledge of acoustic-phonetics. The sudden release of the acoustic pressure introduces a relatively high-energy signal and it is marked as burst onset points and the onset of the first glottal cycle is marked as vowel onset point for non-CLP speakers. In the case of misarticulated phonemes uttered by speakers with CLP, if frication or burst is absent, then the onset of the first formant (F_1) is considered as the start of the vowel. The manually labeled speech events and the associated speech segments are used for training the transformation models to acquire the desired speech template to improve the CLP speech. The manually labeled speech events are also considered as the ground truth for evaluation of the performance of the automatic detection algorithms.

3.3 CLP speech disorder assessment

The recorded speech samples of all the CLP individuals are perceptually evaluated by three well experienced SLPs of AIISH. Each of the expert SLP had an experience of around five years in the field

of CLP speech evaluation. The SLPs assess the phoneme-level intelligibility, word-level intelligibility, articulation, and voice quality by auditory detection and transcription. The perceptual evaluation is conducted in a sound-proof room and all the SLPs used the same computer set-up for listening. During perceptual evaluation, the speech samples are presented to the SLPs in a randomized manner. The SLPs transcribed all the speech samples and provide deviation scores based on the 4−point equal appearing interval (EAI) scale ranging from 0 to 3 [3,8], where 0 =close to normal, 1 =mild deviation, 2 = moderate deviation, and 3 = severe deviation. The higher rating indicates that the misarticulated fricatives/stops are realized as an audible speech deviation. The score zero implies that the CLP speech sample is close to non-CLP fricative/stop sound perceptually.

The study deals with the analysis and modification of misarticulated fricative /s/, misarticulated stops /k/, /t/ and /T/, and nasalized vowels /a/, /i/ and /u/. Therefore, in the following subsections, the assessment of the these speech samples and the corresponding repetitive /CV/ combinations are discussed.

3.3.1 Assessment of misartculated fricative /s/

The fricative /s/ is observed to be one of the frequently occurring misarticulation in the database. It is observed that when the individuals with CLP were expected to utter fricative /s/, some of the speakers tend to replace it with glottal stop or phoneme specific nasal air emission (PSNAE), or palatalized articulation.

Table 3.5: Description of misarticulated fricative data collected from CLP speakers.

Metadata	Category of error	No. of speakers	No. of tokens
/s/	glottal	4	63
	PSNAE	10	119
	palatal	7	57

The fricative /s/ is analyzed in non-meaningful disyllabic /FVFV/ words. In the /FVFV/ word, the F unit corresponds to fricative /s/ and the V unit corresponds to the low vowel /a/ resulting in /sasa/. The corresponding number of speakers and speech tokens considered for misarticulated fricative /s/ are shown in Table 3.5. The speech tokens are selected based on the inter-raters agreement. The SLPs assess the fricative /s/ distortions by transcribing the speech samples and provide deviation

scores on a scale of 0 to 3 as discussed previously. The higher rating indicates that the misarticulated fricative /s/ is realized as an audible articulatory deviation whereas, the score of 0 implies that the misarticulated fricative /s/ is close to non-CLP /s/. Subsequently, using PRAAT software [174], the waveform and spectrographic analysis of the misarticulated fricatives are also performed to decide the category of errors.

Table 3.6: Inter-rater reliability estimation for the misarticulated fricative /s/.

Pair of raters	Cohen's kappa	Kendall's correlation coefficient
1^{st} – 2^{nd}	0.72	0.78
2^{nd} – 3^{rd}	0.69	0.89
1^{st} – 3^{rd}	0.68	0.73

The inter-rater reliability levels are compared using the Cohen's kappa and Kendall's correlation coefficient of concordance. The reliability levels are computed between 1^{st} and 2^{nd}, 1^{st} and 3^{rd} and 2^{nd} and 3^{rd} raters. The correlation coefficient obtained for each pair of raters along with kappa measures are given in Table 3.6. For the fricative /s/ misarticulation, a moderate agreement is observed among the raters with Cohen's kappa value > 0.65. The agreement among the three raters together is assessed using intra-class correlation coefficient (ICC) [175]. Generally, ICC (ρ) is calculated as a ratio of variance of true score to that of total variance. It is given by,

$$\rho = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_v^2} \tag{3.1}$$

where, σ_r^2 and σ_v^2 denote variance of true score among subjects and unwanted variance respectively. σ_r^2 and σ_v^2 constitute the total variance. ICC value describes the correlations within a class of data. In the analysis of misarticulated fricative /s/, an ICC value of 0.8032, within 95% confidence interval $(0.6988, 0.8808)$ is observed. The F value is 13.247 ($p < 0.001$). The ICC value indicate that there is a significant agreement among the raters.

In any case, if the misarticulated fricative segment in a /FVFV/ word is observed to exhibit characteristics like nasal substitution, velar substitution, or any other type of misarticulations, then such segments of the signal are kept unaltered. Since they are likely to illustrate different kind of misarticulations, which are not covered in this work. Therefore, only those /s/ segments are subject to transformations, which acquired one of the above mentioned three types of misarticulations.

3.3.2 Assessment of misarticulated stop consonants

The stop consonants are also analyzed in non-meaningful disyllabic /CVCV/ words. In the /CVCV/ word, the C unit corresponds to one of the unvoiced stop consonants among /k/, /t/, and /T/ and the V unit corresponds to the low vowel /a/ resulting in /kaka/, /tata/ and /TaTa/.

Table 3.7: Description of consonants data collected from CLP speakers.

Metadata	Category of error	No. of speakers	No. of tokens
/k/	glottal	16	298
/t/	glottal	8	140
	velar	9	120
	palatal	3	56
/T/	glottal	9	132
	velar	8	132
	palatal	3	60

The individuals with CLP misarticulate the velar stop /k/ by a glottal stop substitution. For alveolar stop /t/ and /T/, three types of errors are observed in the database, namely, glottal stop substitution, velar substitution and palatalized articulation. A minimum of three repetitions of each of the /CVCV/ words are recorded. The respective number of speakers and tokens considered for the stop consonant analysis and modification are shown in Table 3.7. For the assessment of the three misarticulated stops shown in Table 3.7, the SLPs transcribed the speech samples and provide deviation scores on a scale of 0 to 3 and each scale references are same with that defined in the first paragraph of Section 3.3. Waveform and spectrographic analysis is also performed before the analysis of misarticulated stops using PRAAT software [174] to decide the category of error.

Table 3.8: Inter-rater reliability estimation for the misarticulated stop consonants.

Pair of raters	Cohen's kappa	Kendall's correlation coefficient
1^{st} – 2^{nd}	0.61	0.82
2^{nd} – 3^{rd}	0.74	0.71
1^{st} – 3^{rd}	0.76	0.87

The correlation coefficient obtained for each pair of raters (1^{st} and 2^{nd}, 1^{st} and 3^{rd}, and 2^{nd} and 3^{rd})

along with kappa measures are depicted in Table 3.8. For the misarticulated stops, a moderate agreement is observed among the raters with Cohen's kappa value > 0.7. The agreement among the three raters is assessed using ICC given in equation 3.1 and observed to be 0.7832 with 95% confidence interval $(0.6154, 0.8408)$. The F value is observed to be 14.321 ($p < 0.001$).

3.3.3 Assessment of nasalized vowel phonations

For assessing hypernasal speech, it is reported that test words must comprise of one vowel and one type of pressure consonants. Therefore, for the analysis of hypernasality, consonant-vowel (CV) words are considered in the study. Three repetitions of each of the vowels produced in /CV/ structures (C unit corresponds to /p/) are recorded corresponding to the consonant. Three expert SLPs evaluate the hypernasal speech based on the 4-point rating scale, which reflects increasing hypernasal severity level from 0 through 3. A rating of 0 means that the speech is easy to understand and considered within normal limits, a rating of 1 is considered as mild hypernasality, where the speech is occasionally hard to understand, rating of 2 is considered as moderate hypernasality, where speech is socially unacceptable, and a rating of 3 means severe hypernasality, where the speech is hard to understand most of the time.

Table 3.9: Description of vowel phonation data collected from hypernasal (HN) speakers.

Metadata	No. of speakers	No. of tokens
/a/	39	102
/i/	39	111
/u/	39	98

Table 3.10: Inter-rater reliability estimation for nasalization.

Pair of raters	Cohen's kappa	Kendall's correlation coefficient
1^{st} – 2^{nd}	0.78	0.82
2^{nd} – 3^{rd}	0.73	0.80
1^{st} – 3^{rd}	0.81	0.92

The total number of speakers and tokens included in the study are shown in Table 3.9. The correlation coefficient obtained for each pair of raters (1^{st} and 2^{nd}, 1^{st} and 3^{rd}, and 2^{nd} and 3^{rd}) along

with Cohen's kappa measures are depicted in Table 3.10 for hypernasal speech. For the nasalized vowels, a moderate agreement is observed among the raters with Cohen's kappa value > 0.8. In line with the assessment of misarticulated fricatives and stops, the nasalized vowels are also assessed using ICC, to measure the reliability of the measured data. An ICC value of 0.8132 within 95% confidence interval $(0.7212, 0.8910)$ is observed for hypernasal vowels. The F value of 13.843 ($p < 0.001$) is observed.

3.3.4 Database description for /CVCV/ words

For the combined framework, the repetitive /CV/ combinations of all the phonemes studied in the thesis corresponds to /sasa/, /sisi/, /susu/, /kaka/, /kiki/, /kuku/, /tata/, /titi/, /tutu/, /TaTa/, /TiTi/, and /TuTu/. Accordingly, the total number of tokens used for the mentioned words are shown in the Table 3.11.

Table 3.11: Description of /CVCV/ words collected from CLP speakers.

Category	Glottal stop	Palatal	PSNAE	Velar
/sasa/	63	57	119	-
/sisi/	20	26	62	-
/susu/	29	25	62	-
/kaka/	298	-	-	-
/kiki/	71	-	-	-
/kuku/	70	-	-	-
/tata/	140	56	-	120
/titi/	36	33	-	38
/tutu/	38	33	-	35
/TaTa/	132	60	-	132
/TiTi/	41	36	-	34
/TuTu/	32	36	-	33

Table 3.12: Inter-rater reliability estimation for misarticulated /CVCV/ words.

Pair of raters	Cohen's kappa	Kendall's correlation coefficient
1^{st} – 2^{nd}	0.83	0.89
2^{nd} – 3^{rd}	0.79	0.88
1^{st} – 3^{rd}	0.85	0.94

It is to be noted that the /CVCV/ words comprising of fricative /s/ does not demonstrate velar

substitutions in the database. Similarly, the /CVCV/ words comprising of velar stop /k/ were not uttered correctly and they do not exhibit palatal and PSNAE distortions. In the case of alveolar stop /t/ and retroflex /T/, the speakers does not exhibit PSNAE distortions. The correlation coefficient obtained for each pair of raters (1^{st} and 2^{nd}, 1^{st} and 3^{rd}, and 2^{nd} and 3^{rd}) along with kappa measures are depicted in Table 3.12 for the misarticulated words. For the /CVCV/ words, a moderate agreement is observed among the raters with Cohen's kappa value > 0.7. An ICC value of 0.823 within 95% confidence interval $(0.7011, 0.8990)$ is observed for hypernasal vowels. The F value of 13.532 ($p < 0.001$) is observed.

3.4 Normal speech data

The non-CLP children data is also recorded in the same recording environment, for a minimum of $2-3$ sessions from each speaker. While selecting the non-CLP speakers, care has been taken that they do not have any developmental errors and have adequate hearing and language capabilities. Forty-two

Table 3.13: Description of obstruent data collected from non-CLP speakers.

Metadata	No. of speakers	No. of tokens
/sasa/	30	300
/sisi/	30	102
/susu/	30	102
/kaka/	31	370
/kiki/	31	120
/kuku/	31	120
/tata/	29	336
/titi/	29	135
/tutu/	29	135
/TaTa/	31	355
/TiTi/	31	120
/TuTu/	31	120

non-CLP speakers (22 females, 20 males) in the age range of $8-12$ years have participated in the study. The speakers were asked to repeat the utterances spoken by the instructors and their responses are recorded. The speech data is recorded for the same metadata categories which were used for

the CLP speakers. The total number of speakers and tokens corresponding to non-CLP speakers are shown in the Table 3.13 for obstruents (/s/, /k/, /t/ and /T/) in the vowel contexts /a/, /i/, and /u/ respectively.

Table 3.14: Description of vowel phonation data collected from non-CLP speakers.

Metadata	No. of speakers	No. of tokens
/a/	39	117
/i/	39	117
/u/	39	117

As shown in Table 3.14, for vowel phonations in /CV/ structures (C unit corresponds to /p/), 39 speakers data is considered in the study with 117 tokens for each /CV/ words. In certain cases, non-CLP speech data is used to get insight about the spectral and temporal characteristics of the phonemes, which is helpful in creating a representative template of the general speech. Compared to CLP speakers, the number of tokens for stop consonants and fricatives are much higher for non-CLP speakers. For the combined framework study, the description shown in Table 3.13 is considered.

The non-CLP speakers' speech is used to train the spectral conversion system and to test the performance of a system using objective and subjective measures. During the target selection, age and gender matched CLP and target speaker pairs are considered. In certain cases, if the age and gender of a CLP speaker is similar to more than one target speaker. Then, additional information are taken into account, namely, visual analysis of the prominent spectral concentration, perceptual cues embedded in the consonant-vowel transition regions. Based on the analysis, the non-CLP speaker having the closest desired attributes is chosen as a target. The length of the source speech file is normalized to match with that of the target speech file using the dynamic time warping. Further, the conversion function is trained for each of the respective speaker pairs.

3.5 Summary

This chapter presented the procedure for database development that constitute of both the CLP and non-CLP speech. In this chapter, the details of the assessment of different types of CLP speech disorders are described. Based on the agreement of SLPs, the speech samples are used for the analysis

and modification. Additionally, using PRAAT software, the waveform and spectrographic analysis are performed to characterize the speech disorders in CLP speech. Further, the intra-class correlation and Kendall's correlation coefficient are also computed to objectively compute the rater agreement.

4

Modification of Misarticulated Fricative /s/

Publications

[176] Protima Nomo Sudro, S. R. Mahadeva Prasanna,"Modification of misarticulated fricative /s/ in cleft lip and palate speech", *Biomedical Signal Processing and Control* 67, (2021): 102088.

[177] Sishir Kalita, Protima Nomo Sudro, S. R. Mahadeva Prasanna, S. Dandapat, "Nasal Air Emission in Sibilant Fricatives of Cleft Lip and Palate Speech", in Proceedings of *Interspeech* 2019, pp.4544-4548.

[178] Protima Nomo Sudro, Sishir Kalita, S. R. Mahadeva Prasanna,"Processing Transition Regions of Glottal Stop Substituted /S/ for Intelligibility Enhancement of Cleft Palate Speech", in Proceedings of *Interspeech* 2018, pp.1536-1540.

[179] Protima Nomo Sudro, S. R. Mahadeva Prasanna,"Intelligibility Enhancement of Alveolar Fricative in Cleft Lip and Palate Speech", in Proceedings of *INDICON* 2017, pp.1-6.

Contents

Overview

This chapter describes modification of three types of misarticulated fricative /s/ in cleft lip and palate (CLP) speech. It include palatalized /s/, phoneme-specific nasal air emission distorted /s/, and glottal stop substituted /s/. By using the knowledge of the glottal activity, frication, and silence, an approach is proposed for misarticulated fricative detection and categorization of the type of error. Based on the category of error, an appropriate modification technique is applied. The deviations of palatalized /s/ and phoneme-specific nasal air emission distorted /s/ are corrected by modifying the spectral energy. The high-frequency energy levels are emphasized to improve the perception of fricative /s/ using spectral tilt modification. Glottal stop substitution is modified by inserting artificially synthesized /s/, where the fricative /s/ signal is synthesized using white noise signal and linear prediction filter obtained from non-CLP (healthy) children fricative /s/. Further, the modified speech samples are evaluated using objective and subjective approaches. The evaluation scores obtained from experimental evaluation indicate speech intelligibility improvement of the modified signals compared to the misarticulated fricative /s/.

4.1 Introduction

For non-CLP speakers, the process of generating fricative /s/ through a narrow constriction and creating turbulence in the flow of air requires a complex movement of the articulators [180]. With constriction in the front cavity region mostly, fricative represent a class of sound characterized by high frequency energy. In the case of pathological speakers, the complex articulator configuration makes fricative prone to misarticulations as they require to maintain adequate intra-oral pressure while occluding the impairment. In fricative /s/ misarticulation, deviant power spectra is observed due to shift in spectral prominence, weak frication energy, absence of frication, and additional turbulence noise source created in the nasal cavity. All these factors lowers the required intra-oral pressure for producing fricative sound resulting in misarticulated fricative /s/. Misarticulated fricatives are produced by shifting the place of articulation (PoA) from the front cavity region to the back cavity region of the articulatory system. Thus, shifting the concentration of the fricative energy from the higher to the lower frequency region alters the fricative characteristics. This poses the importance of processing different frequency bands for improving the fricative /s/ misarticulation.

The studies related to misarticulated fricative /s/ modification are quite limited. The exceptions

are two recent works, which include modification of glottal stop substituted /s/ and palatal substitution of /s/ in CLP speech [178,179]. In these studies, acoustic transformations are applied on manually annotated fricative /s/ segments (which requires considerable time and labor). These studies are not feasible for phoneme-specific real-time modifications and it addressed only one type of misarticulation. The intelligibility evaluation of the above-mentioned studies shows significant improvement in speech intelligibility, but phoneme-specific real-time modifications for CLP speech have not yet been attempted. Taking insights from the speech of hearing impaired individuals, where they exhibit high frequency hearing loss. For such cases, studies explored that shifting the higher spectral information into the lower region can improve the intelligibility to some extent. In other studies also, researchers showed that the spectral energy can be either minimized or emphasized based on the desired speech signal characteristics. Hence, similar strategies can be adopted for improving the misarticulated fricative /s/.

The remaining of the chapter is organized as follows: the contribution of the chapter is presented in Section 4.2 and analysis of the misarticulated fricatives is described in Section 4.3. The transformation of misarticulated fricatives, which include automatic segmentation of the fricative errors followed by modification techniques are illustrated in Section 4.4. The impact of modified misarticulated fricative /s/ in intelligibility are explained in Section 4.5. The intelligibility assessments and summary are presented in Section 4.5.1, Section 4.5.2 and 4.6 respectively.

4.2 Contributions

The aim of the work is to investigate the acoustic characteristics of misarticulated fricative /s/ for automatic segmentation, followed by acoustic modification of these errors close to non-CLP like /s/. The detailed description of the speech data used in this chapter are presented in Subsection 3.3.1 of Chapter 3. All the details including the number of speakers, tokens, and the procedure of assessment of the speech disorders by the SLPs were also presented. Three categories of misarticulated fricative /s/ are considered, namely palatalized /s/, phoneme-specific nasal air emission (PSNAE) distorted /s/ and /s/ substituted by glottal stop /ʔ/. For assessing obstruent production errors, in the literature, it is noted that the speech samples must comprise of the target consonant in more than one position per word [8]. Hence, the misarticulated fricative /s/ is studied in the initial and medial position in the fricative-vowel-fricative-vowel (FVFV) structure. Before modification, the automatic segmentation of

the misarticulated fricative /s/ is first performed using the knowledge of the onset of glottal activity, and it is considered an anchoring point. Within 200 ms of the onset point of the glottal activity region, fricative evidence is investigated using band energy ratio and spectral tilt. If fricative is detected, then the category of fricative error is further determined using the features, namely, spectral centroid, dominant spectral centroid, and maximum normalized spectral slope. The detected misarticulated fricative is subject to modification by applying spectral energy compression followed by spectral tilt modification. On the other hand, if no frication is found, then the search interval is analyzed using short-time energy and abrupt transition characteristics. If the region is identified as silent with steep transitions, it is considered glottal stop substituted /s/. Therefore, the glottal stop substituted /s/ is modified by inserting artificially synthetic /s/. The significant contributions of this investigation are stated as follows:

- Analyzing the acoustic characteristics of palatalized /s/, PSNAE distorted /s/, and glottal stop substituted /s/.
- Proposed strategies for the correction of misarticulated fricative /s/:
  - At first, the fricative segments are detected automatically followed by categorization of the type of error into palatalized /s/, PSNAE distorted /s/ and glottal stop substituted /s/.
  - Palatalized /s/ and PSNAE distorted /s/ are distortions in place of articulation (PoA). Hence, spectral modification based approaches are used to correct these errors.
  - A glottal stop substituted /s/ is a distortion in PoA and manner of articulation (MoA) both. Therefore, it is corrected by inserting artificially synthesized /s/.

4.3 Analysis of the misarticulated fricatives in CLP speech

The fricatives represent a class of sound characterized by their manner of production and maximum energy concentrated in the higher frequency region. Fricatives are excited by an aperiodic signal, shaped spectrally by the area and place of narrow supraglottal constriction [181]. The spectral characteristics of fricatives change with the presence and type of obstacle as well as front cavity length downstream of the constriction. Specifically, an alveolar fricative /s/ is a voiceless fricative with the place of constriction in the alveolar ridge. Fricative /s/ in non-CLP speech is acquired through a narrow constriction by creating turbulence in the air flow. The narrow constriction is formed by the hard palate and tongue blade [182]. An adequate intra-oral pressure is required during the /s/ sound

acquisition process [181, 183]. Spectrally, fricative /s/ is characterized by the concentration of energy in the high-frequency region (> 4 kHz). However, in the case of CLP speech, the complex articulator configuration and a lack of adequate intra-oral pressure results in misarticulated fricative [182].

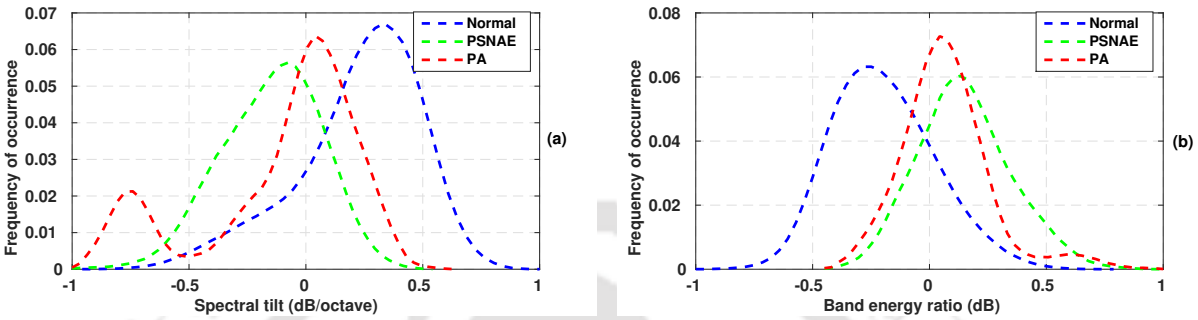

Figure 4.1: Distribution of (a) spectral tilt (dB/octave) and (b) band energy ratio (BER) in dB derived from FFT spectrum for non-CLP /s/, palatalized /s/ (PA) and PSNAE distorted /s/.

The difference between the misarticulated fricative /s/ and non-CLP /s/ are analyzed briefly in the following subsections using spectral tilt and band energy ratio (BER). The spectral tilt is obtained by modeling the spectra using a first-order linear prediction (LP) analysis [184]. The BER is defined as the ratio of spectral energies of two frequency bands [185]. The first band (Eb1) ranges from $[0-4]$ kHz, and another band (Eb2) ranges from $[4-8]$ kHz. BER is given by, $BER = 10 * log10\ \left(\frac{Eb1}{Eb2}\right)$.

4.3.1 Palatalized /s/

The palatalization of /s/ leads to a change in PoA from the alveolar to the palatal region, causing the front cavity to lengthen. As a result, the high-frequency energy shifts towards the low-frequency range [181]. Subsequently, the spectral characteristics of palatalized /s/ exhibit a prominent spectral peak around $[2-4]$ kHz. A study reported that sibilant errors are characterized by the substitution of palatal fricatives for alveolar fricatives [3, 186]. In another study, tongue backing movement is analyzed based on the perceptual and video-fluoroscopic analyses [187]. The palatalized /s/ has different power spectra compared to non-CLP /s/ and palatal fricative /sh/ spectrum [188]. The palatalization of /s/ affects speech intelligibility when formulated in meaningful words or sentences. For example, the words /sasi/ and /shashi/ in the Kannada language corresponds to “plant” and “moon” respectively, which implies that palatalization of alveolar fricative /s/ changes the intelligibility of the word. Therefore, it is important to study the intelligibility of palatalized /s/ in CLP speech.

The distribution of spectral tilt for the palatalized /s/ is shown in Figure 4.1 (a), which includes

both positive and negative values. From Figure 4.1 (a) it is noted that the spectral tilt of non-CLP /s/ have stronger positive slope compared to palatalized /s/. It conveys that the spectral energy of palatalized /s/ is concentrated in the frequency range relatively lower than non-CLP /s/. The additional peak in Figure 4.1 (a) is attributed to the resonances of the back cavity acquiring low amplitude value because of close association with zeros. The posterior movement of the place of constriction drives the speaker to generate a pressure of the fricative before the air is lost through the orifice. Individuals with CLP acquire such misarticulations due to abnormal lingual-palatal contact. It causes a change in the vocal tract volume and front cavity length leading to lower frequency spectral density. From the distribution of BER shown in Figure 4.1 (b), it is observed that palatalized /s/ have higher BER values compared to non-CLP /s/. The dominant high-frequency energy of non-CLP /s/ results in low BER values, and it is reflected in Figure 4.1 (b).

4.3.2 PSNAE distorted /s/

The PSNAE distorted /s/ consists of a turbulence noise source created in the nasal cavity. The turbulence noise produced in the nasal cavity is exhaled forcibly. This noise becomes a part of the generated speech signal, which influences the perceptivity of the listeners [2]. In the case of fricative /s/ distorted by PSNAE, it was reported that associated spectral energy occurs in the frequency range of $[2.5 - 7.0]$ kHz [189]. Because of PSNAE, intra-oral pressure drops, and consequently, fricative energy gets weakened. Fricatives are observed to have a high number of occurrences of audible nasal emission [177, 190]. It is reported that audible nasal emission can reduce intelligibility due to altered acoustic properties [191].

The distribution of spectral tilt is depicted in Figure 4.1 (a) for PSNAE distorted /s/ and it shows strong negative slope relative to palatalized /s/ and non-CLP /s/. It indicates that maximum spectral energy is concentrated in the lower frequency range with energy tailing off toward higher frequencies. From the BER distribution in Figure 4.1 (b), it is observed that PSNAE distorted /s/ has weak frication energy relative to non-CLP /s/. The weak frication energy is due to inadequate intra-oral pressure created during the production of fricative /s/.

4.3.3 Glottal stop substituted /s/

The glottal stop substituted /s/ is produced by tight adduction of vocal folds, thereby accumulating the subglottic air pressure. Then, a sudden release of airflow occurs through an abrupt separation

of the vocal folds. The glottal stop has a severe impact on the intelligibility of CLP speech. A glottal stop may be perceived as a brief choking or popping sound in the throat [192]. In literature, the occurrence rate of the glottal stop in the individuals with CLP is approximately reported to be $60 - 90$ % [193, 194].

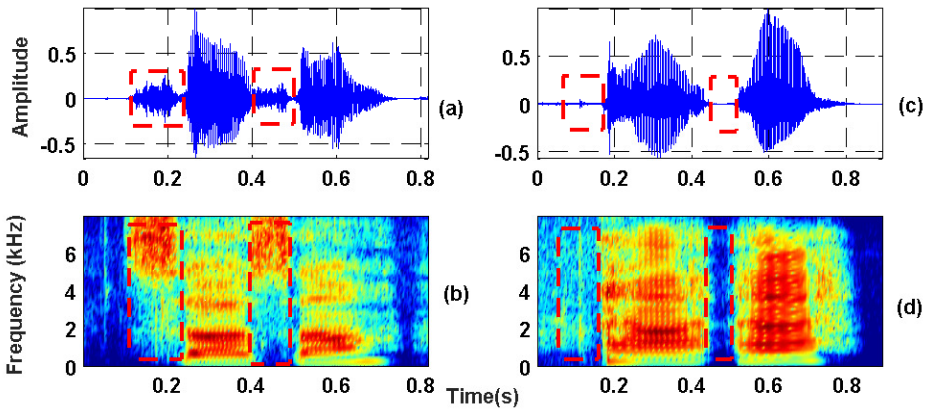

Figure 4.2: Illustration of the waveform and respective spectrogram of a fricative in the intervocalic context /*sasa*/ of a non-CLP speaker (a)-(b) and CLP speaker (c)-(d) with glottal stop substituted /s/.

The silence region marked with a red rectangle in Figure 4.2 (c) and (d) corresponds to the glottal stop substituted /s/ in /sasa/ word opposed to the noise-like frication and high-frequency spectral energy marked with a red rectangle in Figure 4.2 (a) and Figure 4.2 (b) respectively. The absence of formant transitions entering/leaving the adjacent vowel, abrupt start/end of vowel's magnitude, and identifiable stop closure region implies that no constriction has occurred in the oral cavity [195, 196]. Because of the presence of silence region in the glottal stop substitution, BER and spectral tilt are not computed. The average duration of non-CLP fricative /s/ is computed as 134 ms. Therefore, for the glottal stop substituted /s/ in word-initial condition, only 134 ms long duration signal is considered. Hence, in Figure 4.2 (c) and (d) only 134 ms is segmented as glottal stop substituted /s/ besides 190 ms of silence region for word initial condition.

4.4 Transformation of misarticulated fricatives in CLP speech

Based on the analysis of spectral tilt and BER, waveform, and spectrograms depicted in Figure 4.1 and Figure 4.2, the following strategies are developed for the correction of misarticulated fricative /s/.

- Palatalized /s/ and PSNAE distorted /s/ are the distortions in PoA which acquire frication char-

acteristics with deviated spectral characteristics. Hence, spectral modification based approaches are employed for the correction of these errors.

- A glottal stop is characterized by the presence of silence interval and abrupt transitions. Therefore, the modification of silence region is carried out by inserting artificially synthesized fricative /s/.
- Before spectral modification or insertion-based correction, the regions to be modified are segmented automatically. The error category, i.e., frication or silence, is also determined for applying appropriate modification techniques.

4.4.1 Segmentation of misarticulated fricative /s/

The segmentation and classification of misarticulated fricative /s/ utilize the knowledge of the glottal activity, frication, silence, and abrupt transition characteristics. The different stages involved in the segmentation process are described as follows:

Detection of anchor points

In this work, the /FVFV/ words considered have a bisyllabic structure. Each syllable contains the misarticulated fricative, followed by a vowel. The vowels are produced with glottal vibrations, while misarticulated fricatives are produced with the absence of glottal vibrations. As the onset of glottal activity bifurcates the vowel region from the misarticulated fricative region, it is considered an anchor point for the segmentation of misarticulated fricative regions. Accordingly, at first the glottal activity regions are detected using a zero frequency filtering (ZFF) approach [197]. The ZFF process involves the passing of differenced speech signal through a cascade of two ideal zero Hz resonator. The resonator's output contains cumulative DC bias, which is removed by the process of local mean subtraction. The local mean subtracted signal is known as a zero frequency filtered signal (ZFFS). Each of the positive zero crossings of the ZFFS corresponds to the glottal closure instants/epoch locations. The first order slope of ZFFS calculated at each epoch location is termed as the strength of excitation (SoE). Figure 4.4 (b) and (g) depicts that the SoE is comparatively higher for voiced regions relative to unvoiced or silence region. This is because it captures the strength of the quasi-periodic impulse like excitations. Using an appropriate threshold on SoE, the regions with SoE higher than the threshold values are considered as glottal activity regions.

After that, within the glottal activity regions, the syllable nuclei are detected by the method described in [198]. Here, the signal is first band-pass filtered with pass-band frequencies from [0.5 –

4] kHz to enhance the contrast between a vowel and a misarticulated fricative. A windowed frame of 20 ms is considered around each epoch to compute the epoch-synchronous short-time energy of the band-pass filtered signal. The short-time energy contour of the band-pass filtered signal is smoothed using a 100 ms hamming window. The peaks of the smoothed band-pass filtered signal energy profile are used to locate the syllable nuclei. Figure 4.4 (c) and (h) depict the smoothed short-time energy contour of the band-pass filtered signal, and the respective peaks locate the syllable nuclei. Using the information of syllable nuclei location, the anchor point is determined, which primarily denotes the onset of glottal activity region. If the onset point of the glottal activity region lies within the range of 150 ms from the syllable nuclei location, it is considered the anchor point. The anchor points for palatalized /s/ (0.158 s and 0.42 s) and glottal stop substituted /s/ (0.15 s and 0.49 s) are depicted in Figure 4.4 (d) and (i), respectively. Once the anchor points are identified, then the region corresponding to 200 ms before the anchor point is considered as the search interval. Further, this region is processed for the classification of frication versus silence.

Silence versus frication classification

The considered region (200 ms) before the anchor point, may contain silence as in case of glottal stops, while frication in case of palatalized and PSNAE distorted fricative. Therefore, to discriminate the glottal stops from palatalized and PSNAE distorted fricatives, a rule-based approach is employed. From Figure 4.1 (a) and (b), BER and spectral tilt are found to discriminate the fricative errors quite well. Thus, these two features are used to detect the fricative evidence in the /FVFV/ word. For both the quantities, BER denoted by ν_b and spectral tilt denoted by ν_s, separate thresholds (T_b and T_s) are determined experimentally. Using the thresholds T_b and T_s on the fricative evidence (ν), a binary decision is made for the presence/absence of frication as,

$$d_f(n) = \begin{cases} 1 & if \;\; \nu_b[n] > T_b \;\; || \;\; \nu_s[n] > T_s, \\ 0 & otherwise. \end{cases} \tag{4.1}$$

where $T_b = k1 \times \mu_{\nu_b}$ and $T_s = k2 \times \mu_{\nu_s}$. The values of $k1$ and $k2$ are determined experimentally, which are equal to 0.25 and 0.3, respectively. The variables μ_{ν_b} and μ_{ν_s} denote the mean of the BER and spectral tilt computed by averaging the BER and spectral tilt values for the entire search interval. If any of the fricative evidence exceeds the threshold as indicated by equation 4.1, then frication is considered to be present. However, if none of the quantities exceed the threshold, then it is assumed

to be a silence region.

Categorization of misarticulated fricatives

In the final stage, the type of misarticulation (palatalized /s/, PSNAE distorted /s/ and glottal stop) is detected, which is accomplished through the following conditions. If the fricative characteristics is present (i.e., when $d_f(n) = 1$). Further categorization is performed into palatalized /s/ and PSNAE distorted /s/ by using the features namely, spectral centroid (M1), dominant spectral centroid (DSC), and maximum normalized spectral slope (MNSS) [199, 200].

Table 4.1: M1, DSC, MNSS of non-CLP /s/ (NS), palatalized /s/ (PA), and PSNAE distorted /s/ (PSNAE). M1, DSC, MNSS denotes spectral centroid, dominant spectral centroid, and maximum normalized spectral slope.

Category	M1 ($\mu \pm \sigma$) (Hz)	DSC ($\mu \pm \sigma$) (Hz)	MNSS (μ)
NS	5208±1475	4627±1400	2.6×10^{-3}
PA	3233±1436	2691±1283	2.4×10^{-3}
PSNAE	3018±1170	2447±1102	2.1×10^{-3}

At first, the M1 value is computed because it discriminates the region with maximum spectral energy concentration. From Table 4.1 it is observed that the M1 value is relatively lower for palatalized /s/ and PSNAE distorted /s/. It confirms that the place of constriction shifts from anterior to a posterior position in the oral cavity. Although the M1 value distinguishes the fricatives based on the place of constriction, it varies with relative flatness of the spectrum. For example, palatalized /s/ and PSNAE distorted /s/ have a flat spectrum relative to non-CLP /s/, and its M1 value is also low due to the dominance of spectral energy in the low-frequency region. On the other hand, non-CLP /s/ also have a flat spectrum (relatively low M1 value) with the dominance of spectral energy in the high-frequency region. This leads to a significant overlap between M1 values of non-CLP /s/ and misarticulated fricatives. To capture the localized frequency range of the spectral density of misarticulated fricatives, DSC values are calculated. DSC represents the centroid of the values of the magnitude spectrum above the 80 percentile of the distribution. From Table 4.1, it is observed that with the movement of the place of constriction from anterior to the posterior position, DSC values become smaller compared to M1 values. It is also noted that the spread of DSC values are reduced compared to M1 values. This is because of the localized spectral centroid that removes insignificant

signal samples from the spectrum. However, DSC values of PSNAE distorted /s/ and palatalized /s/ have significant overlap. Thus, another feature is required to distinguish them. Hence, MNSS is exploited to separate palatalized /s/ and PSNAE distorted /s/. The MNSS values are computed as the maximum value of the first difference of the spectrum, which is normalized to the sum of the magnitude values of the spectrum. The overall MNSS value averaged across all the frames is observed to be higher for palatalized /s/ compared to PSNAE distorted /s/, which conveys that palatalized /s/ has a higher slope. Using the M1, DSC and MNSS values tabulated in Table 4.1, the palatalized /s/ and PSNAE distorted /s/ denoted by $s_{pa}(n), s_{psnae}(n)$ are categorized as,

$$s_{pa}(n) = \begin{cases} 1 & if\ (1.8 \leq m_1 \leq 4.6)\ \&\ (1.4 \leq DSC \leq 3.97)\ \&\ (0.0022 \leq MNSS < 0.0025)\ \&\ d_f(n) = 1, \\ 0 & otherwise. \end{cases} \tag{4.2}$$

$$s_{psnae}(n) = \begin{cases} 1 & if\ (1.9 \leq m_1 \leq 4.2)\ \&\ (1.3 \leq DSC \leq 3.5)\ \&\ \ (0.0019 \leq MNSS < 0.0023)\ \&\ d_f(n) = 1, \\ 0 & otherwise. \end{cases} \tag{4.3}$$

In case of the absence of fricative evidence in equation 4.1 (i.e., when $d_f(n) = 0$), the region is then analyzed for glottal stop characteristics. The glottal stop substituted /s/ is analyzed using abrupt transitions between the fricative and vowel and vice versa. The abrupt transitions are observed by computing the rate of FV/VF transitions using the function f_{FV}/f_{VF} for each frame of the utterance. The rate of /FV/ and /VF/ transition for a particular /FVFV/ utterance is given by,

$$\begin{aligned} f_{FV}(n) &= ED(\bar{e}_f, e_n) - ED(\bar{e}_v, e_n) \\ f_{VF}(n) &= ED(\bar{e}_v, e_n) - ED(\bar{e}_f, e_n) \end{aligned} \tag{4.4}$$

where $\bar{e}_f$ and $\bar{e}_v$ are the mean short-time energy (STE) computed by averaging STE values over all the frames of misarticulated fricative and vowels, respectively. The variable e_n denote the STE of n^{th} frame and ED indicates the Euclidean distance between two values. From the figure depicted in Figure 4.3, the slope of the transition region (between fricative and vowel or between vowel and fricative) is noted to be less steep for non-CLP /s/ (depicted in Figure 4.3 (a)-(b)) compared to that in-between glottal stop substituted /s/ and vowel (depicted in Figure 4.3 (c)-(d)). The steep slope of the transition region conveys abrupt amplitude increase and abrupt formant transitions in the preceding and next adjacent vowels. Using the knowledge of abrupt transition characteristics and STE below the threshold value T_g, the region with $d_f(n) = 0$ is decided to contain glottal stop substituted

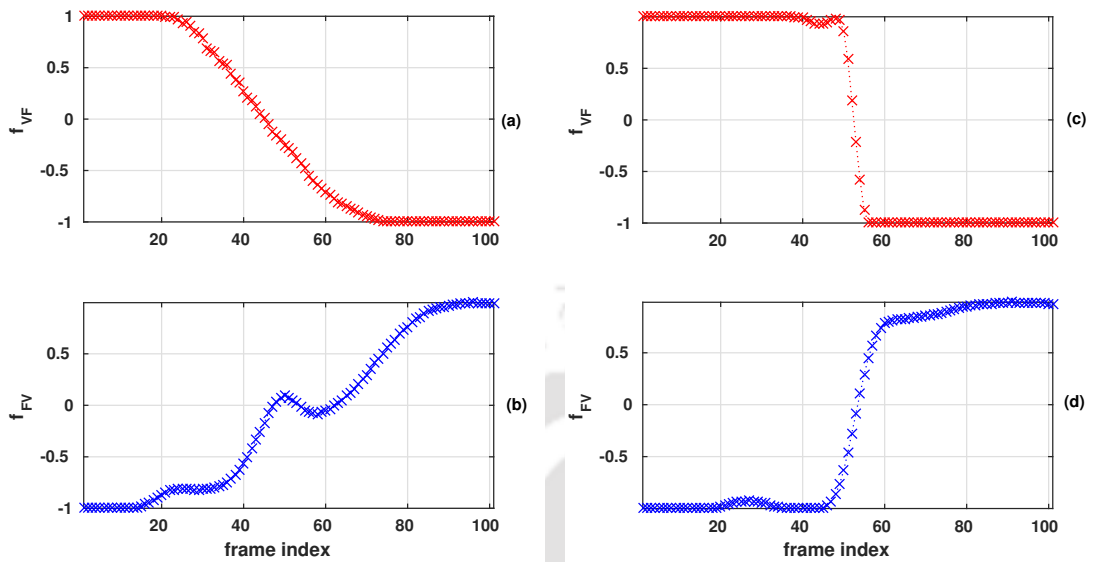

Figure 4.3: Comparison of rate of /VF/ and /FV/ transition region components for, (a)-(b) non-CLP and, (c)-(d) glottal stop substituted CLP speech.

/s/ denoted by $s_{gs}(n)$. Therefore, the glottal stop substituted /s/ is categorized as,

$$s_{gs}(n) = \begin{cases} 1 & if\ STE < T_g\ \&\ f_{FV}/f_{VF} = H\ \&\ d_f(n) = 0 \\ 0 & otherwise. \end{cases} \tag{4.5}$$

where, $T_g = k3 \times \mu_{STE}$ and H denote steep transition slope relative to non-CLP /s/ and vowel transition. The value of $k3$ is determined experimentally, which is equal to 0.02, and μ_{STE} is the mean of the STE computed by averaging the STE values for entire search interval. Therefore, Figure 4.4 (e)-(j) depict the speech waveforms superimposed with segmented fricative errors, namely palatalization of /s/ and glottal stop substituted /s/.

The detection rate of the three above mentioned fricative errors is shown in Table 4.2. The detection

Table 4.2: Performance of misarticulated fricative /s/ detection, palatalization of /s/ (PA), PSNAE distorted /s/ and glottal stop (GS) stop substituted /s/.

Category	PA	PSNAE	GS
Accuracy(%)	84.37	90	88.63

accuracy for palatalization of /s/ is observed to be relatively low compared to PSNAE distorted /s/ and

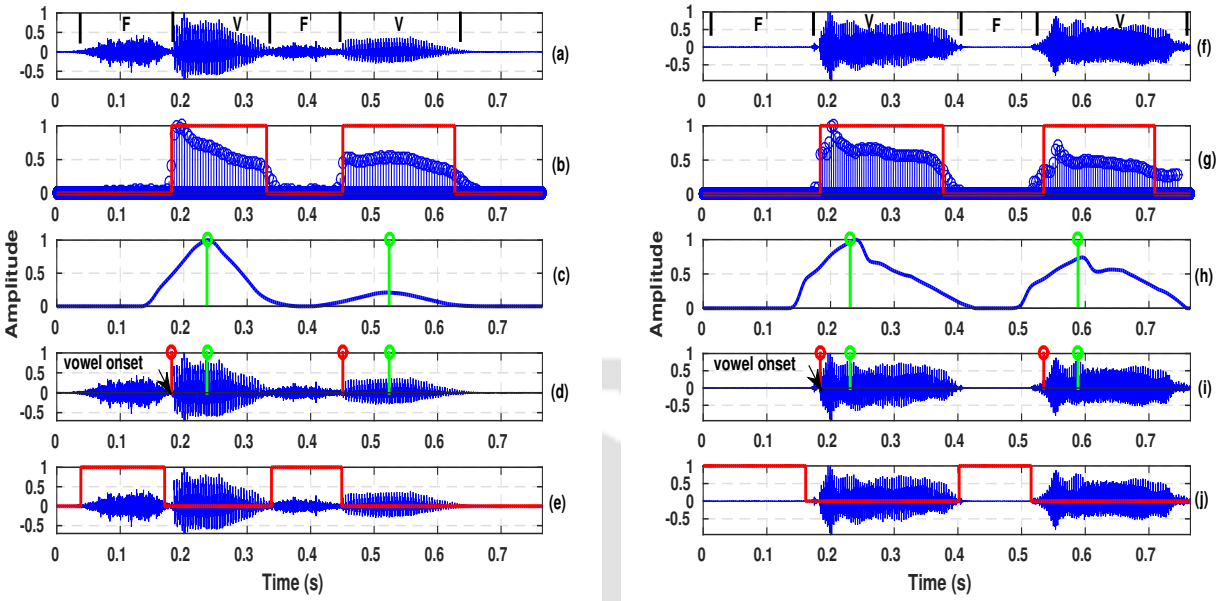

Figure 4.4: Misarticulated fricative segmentation. (a)-(e) and (f)-(j) represent the speech waveforms of /FVFV/ word structure for palatalization of /s/ and glottal stop substituted /s/, SoE superimposed with glottal activity regions, the contour of smoothed STE of band-pass filtered signal with respective peaks representing the syllable nuclei, onset points of glottal activity regions and syllable nuclei locations, and detected fricative errors, respectively.

glottal stop substituted /s/. This is because some of the speech samples of palatalized /s/ show close acoustic characteristics like non-CLP alveolar fricative /s/. However, the characteristics of PSNAE distorted /s/ and glottal stop substituted /s/ are significantly different compared to non-CLP /s/ and hence, results in relatively higher detection accuracy.

4.4.2 Modification of misarticulated fricative /s/

After analyzing the misarticulated fricative /s/ and their corresponding influence in the acoustic characteristics of the signal, the following improvement is performed in this subsection.

Spectral energy compression

From the analysis of misarticulated fricative /s/ in Section 4.3 above, it is observed that palatalized /s/ and PSNAE distorted /s/ retained the frication manner with most of the spectral energy concentrated in the low-frequency range. To modify the misarticulated fricatives similar to non-CLP like /s/, the low-frequency components must be suppressed and emphasize the higher frequency components. Therefore, a sub-band analysis is performed by dividing the fricative spectrum $S(n, \omega_\mu)$ into a lower

and higher frequency band with a cut-off frequency of 4 kHz.

$$S(n,\omega_\mu) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega_\mu m} \tag{4.6}$$

where, $x(m)$ denote the input signal, $w(n)$ is the analysis window, n represents the time at which analysis window is positioned, and ω_μ denote frequency index. The fricative spectrum is estimated from overlapping Gaussian windowed segments (6.4 ms length with 1 ms overlap) and it is analyzed using 2 sub-bands: represented as $S_{\mathrm{LF}}(n,\omega_\mu)$ which ranges from $[0-4]$ kHz and $S_{\mathrm{HF}}(n,\omega_\mu)$, ranging from $[4-8]$ kHz. Here, S_{LF} and S_{HF} denote low and high frequency spectral components, respectively. The spectral energy compression function is applied in the low frequency region $S_{\mathrm{LF}}(n,\omega_\mu)$ of the fricative spectrum. The modified frequency spectrum $\widehat{S}_{\mathrm{LF}}(n,\omega_\mu)$ is given by,

$$\widehat{S}_{\mathrm{LF}}(n,\omega_\mu) = \eta \times (S_{\mathrm{LF}}(n,\omega_\mu)) \tag{4.7}$$

$$where,\ \eta = \sqrt{\frac{\sum_{\mu=\mu_l}^{\mu_u-1}(S_{\mathrm{LF}}(n,\omega_\mu))^2}{\left(\sum_{\mu=\mu_l}^{\mu_u-1} S_{\mathrm{LF}}(n,\omega_\mu) \times \frac{1}{(\mu_u-\mu_l)} \times \sum_{\mu=\mu_l}^{\mu_u-1} S_{\mathrm{LF}}(n,\omega_\mu)\right)^2}}$$

here μ_l and μ_u give the lower and upper boundaries of the frequency interval to be modified. The spectral energy compression factor (η) varies with the total energy in the lower frequency region $[0-4]$ kHz. When the total energy in the S_{LF} region is low, $\eta > 1$ and when energy is high in the S_{LF} region as observed in palatalized /s/ and PSNAE distorted /s/, in that case $\eta < 1$. For the non-CLP speaker's fricative spectrum, η is mostly observed to be greater than 1. Whereas, in palatalized /s/ and PSNAE distorted /s/, η is mostly less than 1. The modified low frequency spectra, $\widehat{S}_{\mathrm{LF}}$ and unprocessed high frequency spectra, S_{HF} are concatenated to get the overall modified spectrum. The modified spectrum is then smoothened by moving averaging filter in order to avoid any sharp transition effect. To further improve the perception of modified /s/ similar to non-CLP /s/, positive spectral tilt modification is performed. Moreover, from Figure 4.1 (a) as well, it is observed that the spectral tilt of palatalized /s/ and PSNAE distorted /s/ are different compared to non-CLP /s/. Therefore, the higher frequency components are emphasized using a positive spectral tilt after spectral energy compression [201].

Insertion method

Another transformation method considered in this study is the insertion method employed for glottal stop substituted /s/ which is produced with a shift in PoA and a change in the MoA as well. The fricative /s/ substituted by the glottal stop is modified by artificially synthesizing fricative /s/ segment similar to the method used for synthesizing stop consonants in [202]. The fricative segment is synthesized using a white noise signal and fricative specific filter obtained from non-CLP /s/. The mean amplitude of non-CLP /s/ is computed as 0.47. For synthesis, a white noise signal is generated with zero mean and unit variance and amplitude not exceeding 0.47. The wide-band spectral content of the white noise signal is passed through a band-pass filter. The center frequency and the cut-off frequencies for the band-pass filter are estimated from the spectrum of the non-CLP /s/. The bandwidth of the band-pass filter is estimated from the frequency at which the spectral energy rises and falls rapidly across the prominent spectral peak [203]. The rising and falling frequency obtained from non-CLP /s/ are averaged across all the speakers. Hence, the lower and the higher cut-off frequencies for the band-pass filter are 3.8 and 7.59 kHz, respectively. Subsequently, the band-pass filtered noise signal is passed through the average linear prediction filter obtained from non-CLP /s/. An LP order of $\frac{fs}{1000}+2$ is used during the computation of average linear prediction coefficients from non-CLP /s/, where fs is the sampling frequency, which is equal to 16 kHz. For word-initial /s/, a fricative signal of 134 ms long duration is synthesized, however, for word medial /s/, the region between the offset of previous phoneme and onset of the preceding phoneme is considered. To observe the perceptual characteristics of synthetic fricative /s/, a perceptual test is performed by replacing the natural /s/ with synthetic /s/ in /FVFV/ words of non-CLP speakers. The listening test showed that 92% (47 out of 51) of the synthetic fricative /s/ sounded similar to non-CLP like fricative /s/.

4.5 Results and discussion

The effectiveness of the transformation method introduced in Section 4.4.2 is examined in this section. The transformation of palatalized /s/ and PSNAE distorted /s/ are shown in Figure 4.5 and Figure 4.6 respectively. For palatalized /s/ depicted in Figure 4.5 (a), it is observed that in the modified spectrum, the prominent spectral density between $[2-4]$ kHz is suppressed compared to the original unprocessed spectrum. However, the higher frequency components are not very high relative to low-frequency components in the unprocessed palatalized spectrum. It indicates that a change

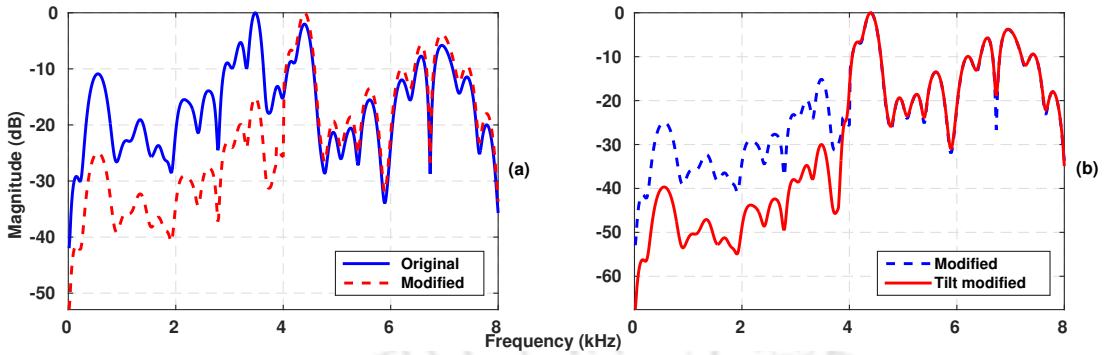

Figure 4.5: Log magnitude spectrum of palatalization of /s/ overlayed with the spectrum of the same after it has been modified for the parameters, (a) spectral energy compression, (b) spectral tilt modification. Original, Modified, and Tilt modified represents the unmodified spectrum, spectral energy compressed spectrum, and spectrum modified by spectral energy compression & spectral tilt transformations.

in spectral energy levels has occurred, but the spectral tilt is not strongly positive, resulting in the low-intensity fricative signal. The positive spectral tilt is important for perceiving the modified /s/ sound like non-CLP /s/ [204]. Consequently, the positive spectral tilt modification is performed on the spectrum after applying spectral energy compression and shown in Figure 4.5 (b). The spectral tilt modified spectrum emphasizes the higher frequency components. Hence, the similarity between non-CLP /s/ spectrum and spectral tilt modified /s/ has increased because it has maximum energy concentrated in the higher frequency region.

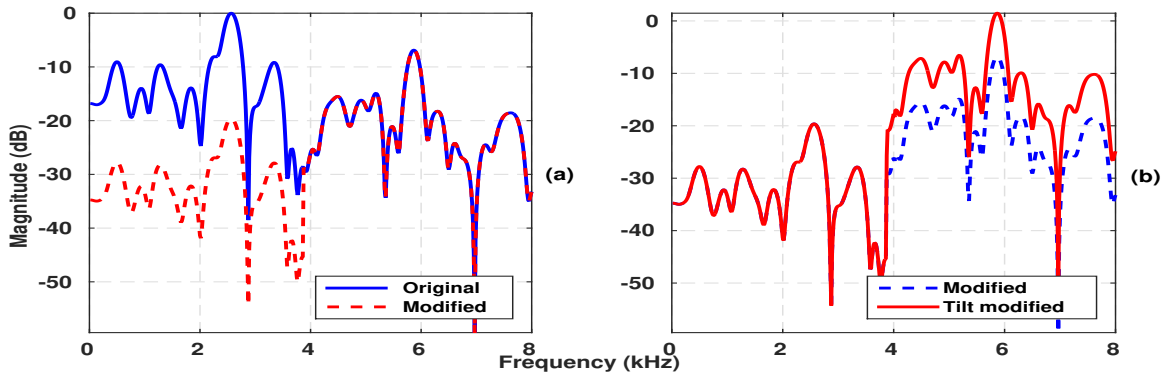

Figure 4.6: Log magnitude spectrum of PSNAE distorted /s/ overlayed with the spectrum of the same after it has been modified for the parameters, (a) spectral energy compression, (b)spectral tilt modification. Original, Modified, and Tilt modified represents the unmodified spectrum, spectral energy compressed spectrum, and spectrum modified by spectral energy compression & spectral tilt transformations.

For PSNAE distorted /s/ shown in Figure 4.6 (a)-(b), consistent observations are noted with re-

spect to palatalized /s/ modifications. In Figure 4.6 (a) the dominant spectral density below 3.7 kHz are attenuated relative to the higher spectral components. The impact of spectral tilt modified spec-

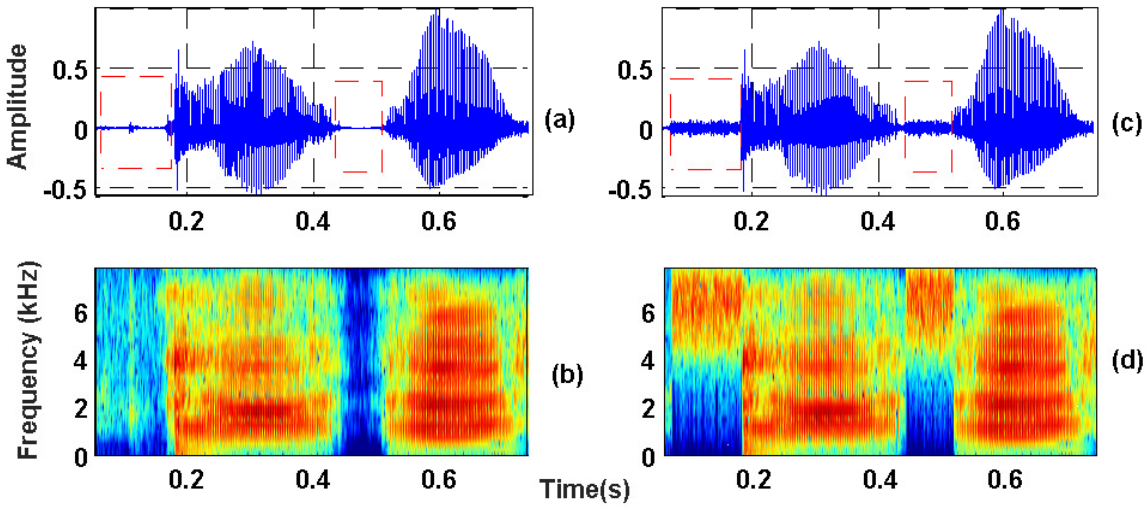

Figure 4.7: Illustration of the unmodified waveform and corresponding spectrogram of a (a)-(b) glottal stop substituted fricative /s/ in the intervocalic context /*sasa*/ and (c)-(d) modified waveform & spectrogram.

trum shown in Figure 4.6 (b) is observed to be higher than the modified spectrum using spectral energy compression alone. After spectral tilt modification, the PSNAE distorted /s/ resembles the spectral characteristics near to non-CLP like /s/ spectrum. The distinguishing feature of fricative /s/ is emphasized which yields prominent high-frequency energy above 4 kHz shown in Figure 4.2 (a)-(b),

The insertion method is analyzed for the correction of /s/ when both the PoA and MoA are changed. The artificially synthesized /s/ sound is inserted in the automatically segmented glottal stop region. The temporal envelope and the corresponding spectrogram of inserted fricative /s/ for glottal stop substitution are depicted in Figure 4.7 (c)-(d). In Figure 4.7 (c)-(d) noise-like fricative signal components with a dominance of high-frequency energy above 4 kHz are observed as opposed to silence in Figure 4.7 (a)-(b) (the region corresponding to the red rectangle). After modification of glottal stop substituted /s/ shown in Figure 4.7 (c)-(d), it nearly resembles the acoustic characteristics of non-CLP fricative /s/ depicted in Figure 4.2 (a)-(b). Further, the impact of modified speech signals on speech intelligibility after performing the transformations are assessed using objective and subjective evaluations, which are discussed in Section 4.5.1 and Section 4.5.2 respectively.

4.5.1 Objective evaluation

The proposed methods are objectively evaluated to check the speech intelligibility improvements after modifications. The objective measures used for the evaluation of modified signals are automatic speech recognition (ASR), spectral centroid (M1), mel-cepstral distortion (MCD), and high to low-frequency energy ratio (HLR).

- **Automatic speech recognition (ASR)**: In this work, the KALDI speech recognition toolkit [205] is used to train the system. The system is adapted to non-CLP children speech, where a hidden Markov model-Gaussian mixture model (HMM-GMM) based method is used to train the system. The recognition results of the distorted fricatives are shown in Table 4.3.

Table 4.3: Phone recognition results for normal fricative /s/ (NS), palatalized /s/ (PA), PSNAE distorted /s/ (PSNAE), and glottal stop substituted /s/ (GS).

Category	NS (%)	Other phonemes (%)	Deletion (%)
NS	87.90	12.10	0.00
PA	32.73	65.27	2.00
PSNAE	19.75	77.25	3.00
GS	6.25	82.75	11.00

For the non-CLP fricative /s/, five training and testing sets are prepared. Each set consists of randomly selected 80% of the samples, which are used for training, and the remaining 20% of the samples are used for testing. The speech recognition system is trained and tested for each of the five sets. The accuracy of non-CLP fricative /s/ averaged across all the five testing sets is observed to be 87.90%. From Table 4.3, it is noted that the recognition accuracy of the misarticulated fricatives is very low. Above 65% of the misarticulated fricatives are recognized as other phonemes, and a small percentage of deletion is also observed. Among all the misarticulated fricative /s/, palatalized /s/ is observed to have the highest recognition rate (32.73%) and lowest for glottal stop substituted /s/ (6.25%). The higher recognition accuracy for palatalized /s/ may be attributed to the place of constriction closer to the alveolar fricative /s/.

The modified misarticulated fricatives are presented to the recognition system adapted to non-CLP /s/. The recognition accuracy of the modified misarticulated fricatives is shown in Table 4.4. From Table 4.4, it is observed that the modified misarticulated fricatives achieve a

Table 4.4: Phone recognition results for modified palatalized /s/ (PAm), modified PSNAE distorted /s/ (PSNAEm), and modified glottal stop substituted /s/ (GSm).

Category	NS (%)	Other phonemes (%)	Deletion (%)
PAm	71.38	28.62	0.00
PSNAEm	67.25	31.75	1.00
GSm	68.24	28.76	3.00

higher recognition rate compared to original misarticulated fricatives. A small percentage of deletion is observed for modified PSNAE distorted /s/ and glottal stop substituted /s/.

- **Spectral centroid (M1)**: The spectral centroid is a feature used in literature to discriminate alveolar fricative /s/ [206]. It represents the highest amplitude peak of the FFT spectrum. Hence, it is used as one of the performance metrics to analyze the impact of transformations on fricative /s/ intelligibility.
- **Mel-cepstral distortion (MCD)**: MCD is used as another performance metric for objective evaluation. It is used to measure the spectral distortion between average non-CLP cepstral coefficients and misarticulated fricative's cepstral coefficients. The MCD value is calculated as,

$$\text{MCD}_t(\text{dB}) = \frac{1}{T}\sum_{t=1}^{T} 10 \times log10 \sqrt{2\sum_{l=1}^{L}(m(l) - \widehat{m}(l))^2} \tag{4.8}$$

where, t and l represent the time frame index and cepstral coefficients index, respectively. $m(l)$ and $\widehat{m}(l)$ denote l^{th} non-CLP cepstral coefficient and l^{th} misarticulated fricative's cepstral coefficient. The overall MCD value is obtained by computing the mean of the MCD values across all the frames.

- **High to low-frequency energy ratio (HLR)**: While performing transformations for the misarticulated fricatives, the higher frequency energy is emphasized compared to low-frequency energy. To take this characteristic into account for modified fricative signals, the high to low-frequency energy ratio is computed. The lower frequency energy ranges from $[0.5 - 4]$ kHz, and higher frequency energy ranges from $[4 - 7.5]$ kHz.

The evaluation is carried out for the /FVFV/ words of the misarticulated fricatives, and the corresponding values are depicted in the boxplot shown in Figure 4.8. The M1 value of palatalized

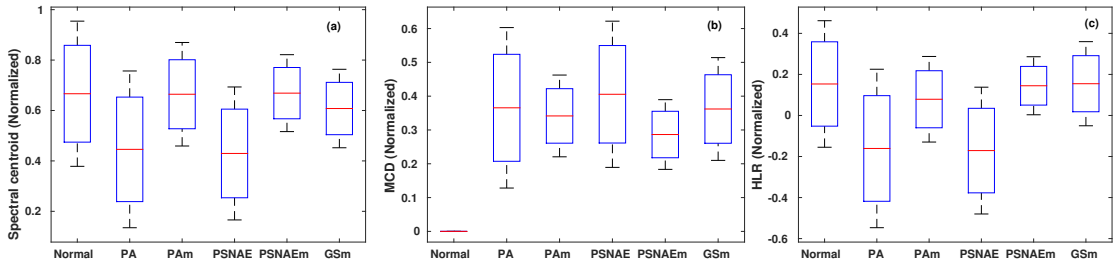

Figure 4.8: Boxplot showing (a) normalized spectral centroid (M1), (b) mel cepstral distortion, and (c) high to low frequency ratio for the phoneme /s/ of non-CLP speech, palatalized articulation, phoneme specific nasal air emission and glottal stop. PA denote palatalized /s/, PAm denote modified palatalized /s/, PSNAE denote phoneme specific nasal air emission distorted /s/ and PSNAEm denote modified PSNAE distorted /s/ and GSm denote modified glottal stop distorted /s/.

/s/ and PSNAE distorted /s/ show significant change with new M1 locations close to non-CLP /s/. This implies that the predominance of spectral energy in the lower frequency region is suppressed and the ones in the higher frequency region are emphasized. Consequently, the M1 value for the corrected glottal stop substitution is slightly lower compared to palatalized /s/ and PSNAE distorted /s/. This conveys that it also possesses M1 in the range similar to non-CLP fricative /s/. The lower MCD values of the modified fricatives indicate that the spectral difference between non-CLP fricative /s/ and misarticulated fricatives are reduced compared to the original unprocessed misarticulated fricatives. The MCD value of non-CLP /s/ is observed to be 0, because the MCD in this work measures the spectral distortion between average non-CLP cepstral coefficients and misarticulated fricative's cepstral coefficients. Furthermore, the HLR values of the modified fricative signals indicate the dominance of higher frequency energy. Thus, the modified speech signals have acquired non-CLP like fricative /s/ characteristics. Modification of the spectral characteristics in case of palatalized /s/ and PSNAE distorted /s/ gives better improvement compared to insertion of artificially synthesized /s/.

A one way Anova test is performed to compare the pre and post modification results. The results for each type of objective measures are tabulated in Table 4.5. Overall, for the spectral centroid an $F = 167.43$ and $p < 0.001$ is observed. Similarly, for mel cepstral distortion, $F = 22.41$, $p < 0.001$ and high to low frequency ratio an $F = 143.96$, $p < 0.001$ are observed respectively. From the p *values*, it is observed that the pre and post modification results are significantly different. Hence, from all the objective metrics, it is observed that the transformation methods give a notable improvement in the

Table 4.5: Results of the one way Anova test for pre and post modified CLP speech. M1 denotes spectral centroid, MCD denotes mel cepstral distortion, and HLR denotes high to low frequency energy ratio. PA denotes palatal articulation, GS refer to glottal stop, and PSNAE denote phoneme specific nasal air emission.

Objective measures	F	p
M1	GS → 1444.29	<0.0001
	PA → 103.61	<0.0001
	PSNAE → 301.34	<0.0001
MCD	GS → 43.53	<0.0001
	PA → 24.16	<0.0001
	PSNAE → 24.65	<0.0001
HLR	GS → 2215.65	<0.0001
	PA → 57.45	<0.0001
	PSNAE → 71.9	<0.0001

modified fricatives.

4.5.2 Subjective evaluation

With the transformations mentioned above, modified signals are subjected to a perceptual test. The listeners might get distracted by other speech-related distortions like the nasalization of vowels along with abnormalities in pitch, loudness, and voice quality, which may affect the intelligibility evaluation of fricative /s/. Therefore, to avoid the naive volunteer listener's distractions, they were given a detailed description of the misarticulations and expected target word before starting the test. A total of 10 naive listeners participated in the perceptual test, and they bear the knowledge of speech technology. All the listening is made through headphones.

Speech quality assessment

The speech quality of the misarticulated fricatives and the modified fricatives are assessed using a 5-point rating scale mean opinion score (MOS) (1 = bad, 2 =fair, 3 = good, 4 = very good, 5 = excellent). The speech samples are randomly numbered to avoid any bias towards the information of the modified signal. For each type of misarticulations, ten modified speech utterances, ten original misarticulated utterances, and ten non-CLP /s/ utterances are presented to the listeners. The distribution of the MOS values, evaluated by naive listeners is shown in Figure 4.9.

From the boxplot depicted in Figure 4.9, it is observed that the MOS of non-CLP /s/ is higher than modified /s/. From Figure 4.9, it is also observed that the glottal substituted /s/ shows a

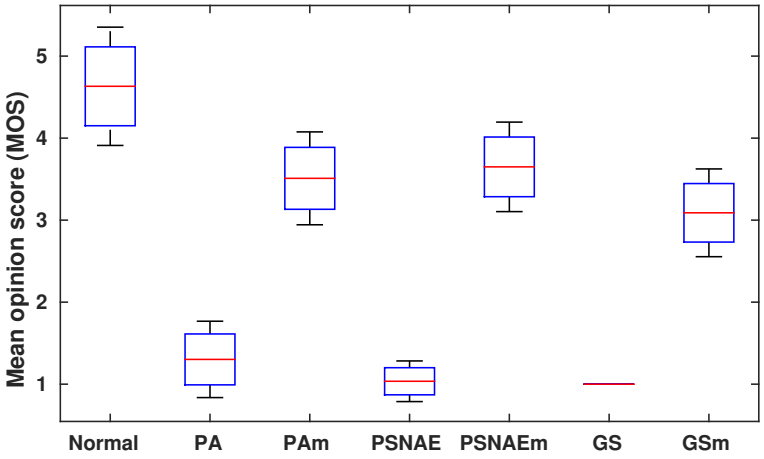

Figure 4.9: Boxplot showing mean opinion scores (MOS) for the phoneme /s/ of non-CLP speech, palatalized articulation, phoneme specific nasal air emission and glottal stop. PA denote palatalized /s/, PAm denote modified palatalized /s/, PSNAE denote phoneme specific nasal air emission distorted /s/ and PSNAEm denote modified PSNAE distorted /s/, GS denote glottal stop substituted /s/ and GSm denoted modified glottal stop distorted /s/.

MOS of 1, and it may be related to the abrupt adduction of vocal folds [192]. Palatalized /s/ and PSNAE distorted /s/ exhibit a MOS below 2. However, the modified misarticulated fricative show MOS greater than 3. This implies that the modified /s/ are preferred over the misarticulated /s/.

Preference test on similarity

The evaluators are asked to give preferences for the speech samples whose intelligibility resembles near to non-CLP fricative /s/ sound. If no improvement in intelligibility is perceived, then they can mark it as "No Preference". Table 4.6 shows the percentage of evaluated speech samples for the original, modified, and no preference conditions.

Table 4.6: Results of comparison test for palatalized /s/ (PA), PSNAE distorted /s/ (PSNAE), and glottal stop substituted /s/ (GS) by naive listeners.

Category	Original (%)	Modified (%)	No Preference (%)	p-value
PA	15.7	83.2	1.11	< 0.001
PSNAE	14.3	77.8	7.9	< 0.003
GS	3.3	90	6.7	< 0.001

The corresponding values for each pair of speech utterances revealed intelligibility improvement

for the modified speech signals. However, the improvement for the case of palatalized /s/ is somewhat less than the other type of modified misarticulations. This is because the PoA of palatalized /s/ may be closer to non-CLP /s/, which reduces the perceptual difference between the two fricatives. Between each pair of the speech utterances, a pairwise t-test is performed to observe the differences between the distributions of perceptual test values obtained from naive listeners. All pairs are observed to be significantly different, with a p-value of < 0.005. The differences are substantially higher in the case of glottal stop as compared to PSNAE distorted /s/ and palatalized /s/ because the high acoustic contrast between glottal stop substituted /s/ initially and high-frequency dominated fricative /s/ after modification affects the perceptivity significantly. As a result, the modified glottal stop misarticulation is perceived differently compared to the lack of frication nature in unprocessed glottal stop substituted /s/. Subjective evaluation results indicate that the listeners preferred to listen to enhanced CLP speech samples modified by the proposed approach over original samples.

From the analysis and experimental evaluation above, it is noted that the proposed approach will automatically detect and modify three misarticulated fricative /s/: palatalized /s/, PSNAE distorted /s/ and glottal stop substituted /s/. It is speculated that the proposed method may work for other fricatives and affricates for the aforementioned three types of errors. However, the study is limited to a specific phoneme and three types of misarticulations only. The study is also restricted to a nonsensical /FVFV/ word. In a realistic scenarios, the segmentation accuracy may vary due to reverberation and background noise, and it may further affect the enhancement.

4.6 Summary

This work performed the analysis, segmentation, and modification of misarticulated fricative /s/ to enhance the CLP speech intelligibility. The proposed algorithm detects the misarticulated fricatives with a significant detection rate. For the misarticulated fricatives which possess deviated fricative /s/, the corresponding spectral characteristics are modified using spectral energy compression and positive spectral tilt modification. Whereas, the misarticulated fricatives which acquire identifiable stop closure region with abrupt transition characteristics are modified via inserting artificially synthesized /s/. To evaluate the effect of exploited modification methods, subjective and objective assessments are performed. Objective tests have confirmed that the modified misarticulated fricative's spectral characteristics are closer to the non-CLP fricative /s/. The results from subjective test convey that

the modified fricative signals have higher intelligibility compared to misarticulated fricatives. From the subjective assessment, it is observed that modified misarticulated fricative /s/ achieves a relatively lower MOS than the non-CLP /s/. This difference in MOS values may be related to the thresholds and modification parameters.

This chapter analyzed and modified only one frequently distorted sound unit, i.e., misarticulated fricative /s/ and it is studied in one vowel context in a nonsensical /FVFV/ word [176]. However, vowel nasalization, and other misarticulated stop consonants also distort CLP speech intelligibility. Therefore, it is also necessary to study and modify consonant misarticulations and vowel nasalization to improve the perceptivity of the CLP speech. In the next chapter, three stop consonant are analyzed and modified corresponding to glottal, velar and palatal substitutions.

5

Event-Based Transformation of Misarticulated Stops

Publications

[207] Protima Nomo Sudro, Vikram C. M, S. R. Mahadeva Prasanna, "Event-Based Transformation of Misarticulated Stops in Cleft Lip and Palate Speech", *Circuits, Systems, and Signal Processing*, 40(8), 2021: 4064-4088.

[208] Protima Nomo Sudro, " Intelligibility Enhancement of Cleft Lip and Palate Speech", 5^{th} *Doctoral consortium, Interspeech* 2019.

Contents

Overview

This chapter focuses on the modification of misarticulations produced for unvoiced stop consonants in cleft lip and palate (CLP) speech. Three types of misarticulations are studied: glottal, palatal, and velar stop substitutions. The stop consonants are misarticulated due to inadequate build-up of intra-oral pressure caused by velopharyngeal dysfunction and oronasal fistula. The misarticulated stops affect the speech intelligibility and quality. The misarticulated stops are analyzed and modified using the speech data collected from Kannada speaking children. An event-based modification approach is used to correct three misarticulated stops, /k/, /t/, and /T/. At first, automatic detection of burst onset and vowel onset events is carried out. Then, the region from vowel onset to 20 ms duration of the vowel is extracted. Further, the region from burst onset point to 20 ms duration of the vowel is defined as the region for modification. It is transformed using the nonnegative matrix factorization method. The objective and subjective evaluation results show that the proposed event-based transformation approach provides a relative improvement compared to the entire-word modification (signal processed without using the knowledge of burst and vowel onset events). The event-based transformed misarticulated stops showed close similarity with the non-CLP (healthy) stop consonants in terms of perceptual quality. The improved performance accuracy of modified stops suggests that the speech distortion is minimized.

5.1 Introduction

Stop consonants are characterized by multiple sub-phonetic units, namely, the onset of closure, closure interval, burst-onset, and voice-onset time [209]. The closure interval denotes a state when the articulators are held together, forming an oral occlusion (closure) behind which pressure is built up. During the closure interval, the vocal folds may or may not vibrate. If the vocal folds vibrate, it is considered as prevoicing, and termed as voiced stop consonants. When the pressure is released suddenly, it introduces a relatively high energy burst signal [210]. The next attribute after the burst release is the VOT, which is characterized by an interval between the onset of the stop-burst and the onset of the vowel [211]. In CLP speech, the stops are misarticulated due to the deviation in voicing (devoicing errors), place of articulation (glottal, velar, palatal stops), and manner of articulation (weak and nasalized stops). Unlike the misarticulated stops in CLP speech, the acoustic characteristics of the sub-phonetic units of stops uttered by non-CLP speakers do not exhibit many variations.

Several studies in the literature reported the occurrence of misarticulated stops in CLP speech. In

this work, compensatory error for three stop consonants in the vowel context /a/ is addressed. The study concentrates on the compensatory errors because it can be corrected using behavioural therapy as compared to the obligatory errors which can be corrected using surgical intervention only. Hence, the studies related to the compensatory errors in CLP speech are briefly reviewed. The perceptual experiments conducted in [212] revealed the presence of mid-dorsum palatal stops in the speech of repaired cleft palate speakers. The electropalatographic experiments conducted in [213] and [214] reported palatalized articulation and velar-alveolar double articulations for the velar and alveolar sounds in speakers with repaired CLP. The study showed that there is a reduction in the contrast between alveolar and velar sounds. In [215], it is reported that compensations such as the reduction in alveolar and velar contrast, palatalization, and velar backing, occur due to the presence of a fistula and dental arch dimensions. A study in [216] carried out the spectral analysis of alveolar (/t/) and velar (/k/) stops produced by children with CLP. It is noted that, due to palatalization or backing errors, there is a change in the first spectral moment of /t/. The difference between the acoustic characteristics of /t/ and /k/ phonemes is also reduced. A study in [217] analyzed the spectral moments of glottal stop substitutions for alveolar stops. All the above studies showed that articulatory impairment distorts the characteristics of stop consonants and this in turn affects the speech intelligibility and quality.

From Chapter 1 and aforementioned studies, it is noted that misarticulated stops affects the CLP speech intelligibility and quality. This provide a room for the researchers to explore possible ways for the enhancement of CLP speech intelligibility.

5.1.1 Challenges

In most of the speech enhancement systems reported in Chapter 2, the entire speech signal is processed, which involves block-processing the signal with fixed window size. The fixed frame size models might not accurately represent speech dynamics. The speech dynamics here denotes the rapid change in temporal and spectral characteristics, for example, the transitions from closure interval to burst-onset, burst onset to voice onset time (VOT), and VOT to the onset of adjacent vowel. It is reported that the different attributes of the speech dynamics characterized the stop consonants [218]. It is also reported that speech dynamics are essential because they carry significant perceptual cues related to the speech intelligibility [218, 219]. If the block-processing of the signal does not comply with the speech dynamics. In that case, the source spectral components might mix into irrelevant target spectral components, thus deteriorating the quality of the enhanced speech [136, 146].

In the case of CLP speech, the presence of articulation errors in unvoiced stops are characterized by the absence of formant transitions, absence of burst, presence of weak burst, and weak spectral prominence [215, 220]. These articulation errors occur due to change in place of articulation (PoA), or manner of articulation, or both in response to articulatory impairment. These factors further affect CLP speech intelligibility. For illustration, the difference between the non-CLP speaker's stop and CLP speaker's misarticulated stop is depicted in Figure 5.1.

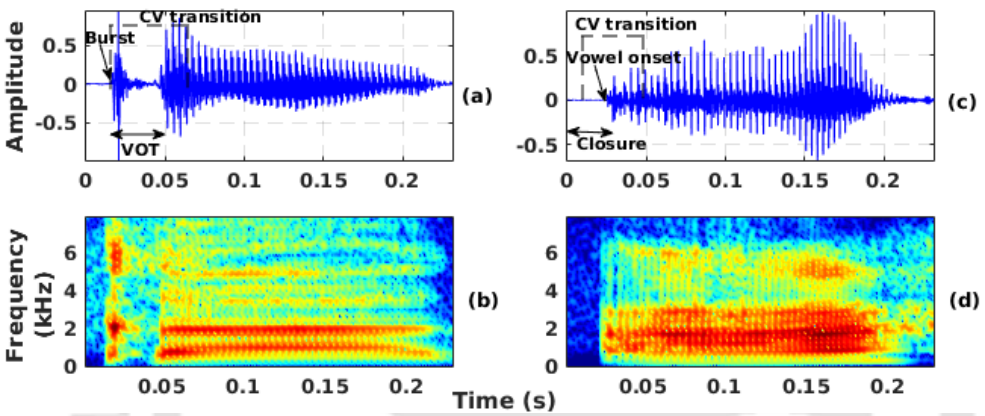

Figure 5.1: Illustration of speech dynamics for (a) CV (C and V correspond to /k/ and /a/, respectively) unit containing stop consonant of non-CLP speaker, (b) the corresponding spectrogram, (c) CV (C and V correspond to /k/ and /a/ respectively) unit containing glottal stop substitution, and (d) the corresponding spectrogram.

Figure 5.1 (a)-(b) shows the waveform and spectrogram of a non-CLP velar stop /k/. In Figure 5.1 (a), a burst signal is observed around 0.02 s, which is reflected as an evident spectral prominence in the spectrogram shown in Figure 5.1 (b). The waveform and spectrogram of a glottal stop produced for the target /k/ by the CLP speaker are depicted in Figure 5.1 (c)-(d), respectively. The glottal stop does not exhibit burst like the non-CLP velar stop /k/, due to the vocal fold tight adduction. From Figure 5.1 (c)-(d), it is also observed that there is an abrupt start of the vowel magnitude with no formant transitions. This implies that no constriction has occurred in the oral cavity. From Figure 5.1 (c)-(d), it is observed that the misarticulated stop and the corresponding transition region are affected due to the articulatory impairment. Hence, it becomes essential to analyze and process the speech dynamics associated with stops for CLP speech enhancement, rather than processing the entire speech signal at a time [146, 221]. The acoustic characteristics of the speech dynamics in CLP speech exhibit different degrees of variations. The speech dynamics of the misarticulated stops can be analyzed by anchoring around the burst onset and vowel onset events [222]. Therefore, it will be

possible to select and apply different analysis methods around different events of the misarticulated stops in CLP speech.

In the literature, it has been indicated that the voice conversion (VC) method is a viable technique for modifying articulation-related parameters of speech [1, 53, 54, 162]. However, speech enhancement using the model-based VC technique requires a large amount of training data. It is difficult to obtain a sufficient amount of pathological speech data for training. With insufficient data, the model-based VC technique might not preserve the dynamic spectral details of the speech sub-phonetic units, and it may incur an over smoothing effect, resulting in muffled speech. To overcome the smoothing problem caused by insufficient training data, non-parametric exemplar-based VC techniques have been proposed in literature [223]. Further, to efficiently estimate the distinct acoustic characteristics of the source and target speech, a joint nonnegative matrix factorization (NMF) framework is proposed in the literature [224]. It showed to improve the enhanced speech quality significantly.

Additionally, the deviant acoustic characteristics of misarticulated stops present a significant challenge in the segmentation of event-locations and the corresponding transitions. Hence, it is essential to incorporate the knowledge of the deviant stop characteristics in the transformation model to achieve an intelligible speech. Therefore, to add dynamic information in the transformation system, event-based approaches are suggested in the literature [146, 221]. In the event-based approach, the speech signal is processed by anchoring it around some specific speech events [225, 226]. Motivated by the previous studies, in this work, an approach is proposed to enhance the CLP speech using the NMF framework and event-based processing.

5.1.2 Contributions

Motivated by the need to correct the articulatory error, an event-based modification of misarticulated stops is performed in this chapter, using an exemplar-based VC technique. The detailed description of the speech data used in this work is presented in Subsection 3.3.2 of Chapter 3. The data description include the variants of speech distortions, total number of speakers, and speech samples considered in the study. The proposed approach consists of the following contributions.

- The articulatory errors are analyzed for the unvoiced stops /k/, /t/, and /T/, corresponding to the glottal, palatal, and velar stop substitutions.
- The speech events corresponding to burst onset point and vowel onset point (VOP) are detected. The region starting from burst onset point to 20 ms duration of the vowel is considered as the

region for modification.

- NMF method is used to learn the transformation model from the CLP and non-CLP speech spectra and further use the learned model for modification.
- The modified signal obtained using the event-based approach is compared with the entire speech signal modification, which is processed without using the event-knowledge.

The rest of the chapter is organized as follows: The analysis of misarticulated stops and segmentation is presented in Section 5.2. The NMF approach for the correction of articulatory errors and its impact is discussed in Section 5.3. The objective and subjective evaluation results are presented in Section 5.4. Finally, Section 5.5 gives the summary and scope for the subsequent chapter.

5.2 Analysis and Segmentation

In this section, the misarticulated stop consonants are analyzed by comparing it with the corresponding non-CLP stop consonants. Figure 5.2 depicts the waveform and spectrogram of /CV/ transition region of a non-CLP /ta/ syllable, CLP /ta/ syllable where the stop /t/ is substituted by: glottal stop, palatalized stop, and velar stop, respectively.

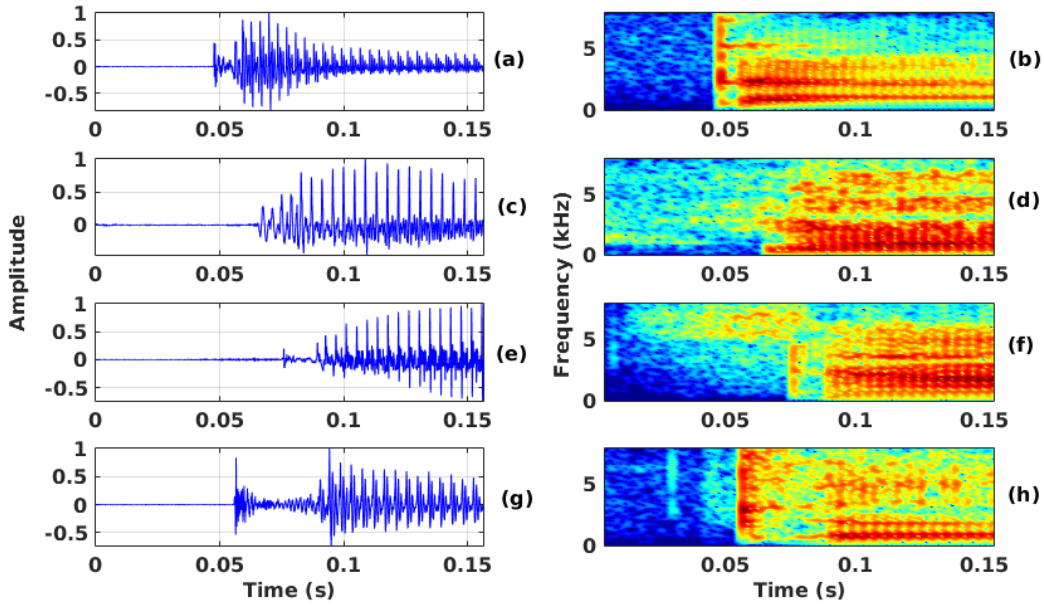

Figure 5.2: Illustration of the waveform and corresponding spectrogram of /ta/ syllable for (a)-(b) non-CLP speaker, (c)-(d) CLP speaker, where /t/ is substituted by glottal stop sound, (e)-(f) CLP speaker, where /t/ is substituted by palatalized stop, and (g)-(h) CLP speaker, where /t/ is substituted by velar stop.

In the case of a non-CLP stop consonant /t/ depicted in Figure 5.2 (a)-(b), the spectral prominence is observed to be ranging from mid to higher frequency region, as the PoA is present in the dental region. However, due to the articulatory impairments in CLP speech, the stop /t/ is substituted by the glottal stop sound, which is shown in Figure 5.2 (c)-(d). Here, the onset of the spectral prominence is observed as that of vowel region. The absence of formant transitions entering/leaving the adjacent vowel, abrupt start/end of vowel magnitude, and identifiable stop closure region implies that no constriction occurred in the oral cavity [195]. When the stop /t/ is uttered as a palatal stop in CLP speech, the burst spectrum deviates from the dental stop /t/. The deviant burst spectrum due to palatal substitution is depicted in Figure 5.2 (e)-(f). Figure 5.2 (f) shows that the spectral prominence shift towards low-frequency regions resulting in palatalized articulation of the stop /t/. The VOT of palatalized /t/ is also longer compared to the dental /t/. When the stop /t/ is substituted by velar stop /k/, a strong spectral prominence is observed around 2.5 kHz. From Figure 5.2 (g)-(h) and Figure 5.2 (a)-(b), it is also observed that the VOT of velar /k/ is longer in duration compared to the VOT of dental /t/.

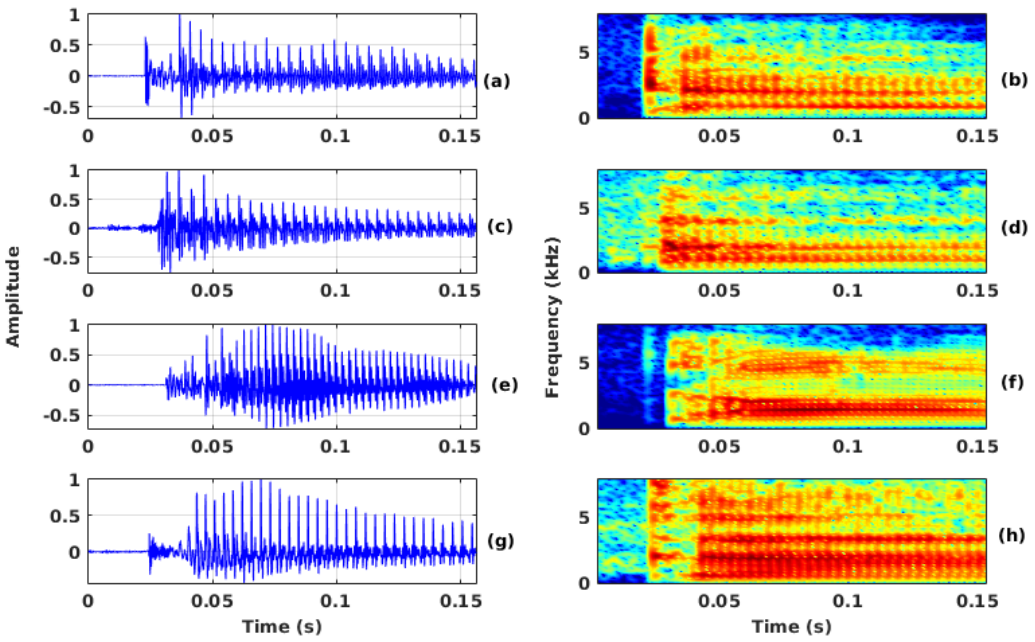

Figure 5.3: Illustration of the waveform and corresponding spectrogram of /Ta/ syllable for (a)-(b) non-CLP speaker, (c)-(d) CLP speaker, where /T/ is substituted by glottal stop sound, (e)-(f) CLP speaker, where /T/ is substituted by palatalized stop, and (g)-(h) CLP speaker, where /T/ is substituted by velar stop.

The characteristics of retroflex /T/ of a non-CLP speaker is depicted in Figure 5.3 (a)-(b), where an evident spectral prominence above 2 kHz characterizes the non-CLP stop /T/. The glottal stop substituted /T/ is shown in Figure 5.3 (c)-(d). Similar to glottal stop substitution for /k/ and /t/, the glottal stop substituted /T/ is also observed to be characterized by the absence of burst and the abrupt start of the vowel magnitude. The palatalized /T/ shown in Figure 5.3 (e)-(f) exhibits a different burst spectrum with prominent frequency concentration below 2.5 kHz and above 5 kHz. In comparison, non-CLP /T/ is characterized by frequency concentration ranging from 2.5 kHz to a higher frequency region. The VOT of palatalized /T/ is longer than the non-CLP stop /T/. The velar stop /k/ substituted for the target /T/ is shown in Figure 5.3 (g)-(h). In the case of velar stop /k/ uttered for the target /T/, prominent spectral energy is observed around 2 kHz as compared to the non-CLP /T/. Here, the spectral energy is concentrated above 2.5 kHz. The VOT is also comparatively longer in velar stop /k/ uttered for the target /T/.

From Figures 5.1, 5.2, and 5.3, it can be noted that the burst and the corresponding speech dynamics tend to deviate for the misarticulated stops. To transform the misarticulated stops, these deviations need to be corrected. In this work, /CVCV/ words are used for the analysis of misarticulated stops. In the /CVCV/ words, it is observed that VOP bifurcates the stops from the vowels. In the case of misarticulated stops in CLP speech, the region at the left side of the VOP contains burst deviation or absence of burst. The region on the right side of VOP contains formant transitions that are affected due to coarticulation. Hence, VOP is considered as the potential anchor point for the segmentation and modification.

5.2.1 Event knowledge-based segmentation

The signal segmentation begins with detecting the glottal activity region, followed by calculating the plosion index (PI) for capturing an abrupt increase in energy. At first, the glottal activity regions are detected using a zero frequency filtering (ZFF) approach [197]. The ZFF process involves passing a differenced speech signal through a cascade of two ideal zero Hz resonators. The resonator's output contains cumulative DC bias, which is removed by local mean subtraction. The local mean subtracted signal is known as a zero frequency filtered signal (ZFFS). Each of the positive zero crossings of the ZFFS corresponds to the glottal closure instants/epoch locations. The term epoch refers to the instant of significant excitation. The first order slope of ZFFS calculated at each epoch location is termed the strength of excitation (SoE). In the /CVCV/ word the glottal activity regions correspond

to the vowel regions. Thus, it is appropriate to search the burst before the vowel regions. Therefore, a search interval of 120 ms corresponding to maximum VOT (measured across the world's different languages) [211] is used to detect the burst. A non-linear temporal measure called PI is used for the detection of burst [210]. PI is used to capture the intrinsic nature of a transient-like signal preceded by a low-level signal, as in a closure-burst transition. PI is defined as the peak amplitude ratio in the transient to the average of absolute values over an appropriate interval, excluding the peak amplitude's immediate neighborhood. PI is defined as

$$PI(m_o, n_1, n_2) = \frac{|X(m_o)|}{X_{avg}(n_1, n_2)} \tag{5.1}$$

$$X_{avg}(n_1, n_2) = \frac{\sum_{i=m_o-(n_1+n_2)}^{i=m_o-(n_1+1)} |X(i)|}{n2} \tag{5.2}$$

where m_o denotes the sample of interest, i.e., the instant of maximum amplitude, and X_{avg} denotes the average absolute amplitudes of n_2 interval. n_1 and n_2 in equation 5.1 are the parameters representing an interval respectively. The n_1 interval corresponds to the low-level noise-like signal components called pre-frication. The pre-frication is usually observed preceding the instant of maximum amplitude within an unvoiced stop-burst. The n_2 interval corresponds to a segment in the stop closure region with amplitudes lower than those in the pre-frication region (n_1). n_1 and n_2 corresponds to an interval ranging for a duration of 6 ms and 16 ms, respectively. Thus, for the 16 kHz sampling frequency, $n_1 = 96$ samples and $n_2 = 256$ samples. n_1 and n_2 values are set based on the experimental results reported by the authors in [210,227] for different datasets consisting of different languages, read speech, and conversational speech. The experiments are repeated for the CLP speech database and observed that setting $n_1 = 6$ ms and $n_2 = 16$ ms results in significant detection accuracy.

In the literature, the threshold of PI corresponds to 9 dB [210, 225, 228], and the same is used for this work. The threshold of PI is selected based on the equal error rate criteria using the TIMIT database. The authors in [210] showed that 9 dB of threshold works well on the TIMIT database under clean and noisy conditions. They conducted burst detection experiments on Indian languages and Buckeye conversational speech database using the same threshold, and report significant accuracy.

For illustration, the detection of burst onset and vowel onset are shown in Figure 5.4. Figure 5.4 (a) shows the speech waveform corresponding to the /CVCV/ word containing a palatalized stop. Figure 5.4 (b) and Figure 5.4 (c) shows the ZFF signal and the SoE, respectively. The detected

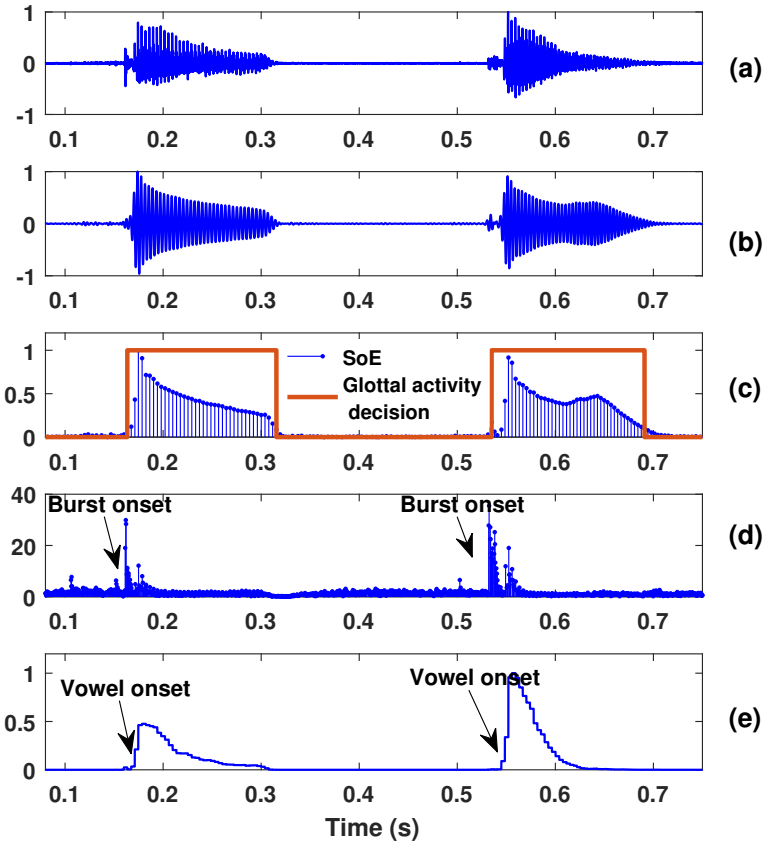

Figure 5.4: (a) Speech waveform corresponding to /CVCV/ word containing palatal stops, (b) zero frequency filtered signal (ZFFS), (c) strength of excitation (SoE) with glottal activity decision, (d) PI with the detected burst, and (e) maximum weighted inner product (MWIP) with the detected VOP.

glottal activity regions corresponding to the regions with SoE higher than the threshold values are also indicated in Figure 5.4 (c). The PI evidence, along with the detected burst onset, is depicted in Figure 5.4(d). PI shows higher values around the burst onset. The burst candidates are detected using the procedure explained in [210]. Along with the genuine burst candidate, there may be some spurious ones getting identified as a burst. Hence, to avoid false detection, the knowledge of the glottal activity region is used. The burst in the C unit of the /CVCV/ word is always present before the vowel. Therefore, the onset of the vowel is first located using glottal activity evidence. Next, within the predefined search interval, a burst candidate detected by the PI method is considered the burst onset location. To avoid the false detection of vowel onsets as a burst, a maximum normalized cross-correlation coefficient is used. It is computed for the speech frames of two adjacent glottal cycles.

The glottal activity onset event is considered as the vowel onset event in the unvoiced stops. Vowel onset and glottal activity onset events are referred to as the same. From the glottal activity onset, 120 ms right, and 120 ms left is considered the search interval for the burst detection.

As shown in Figure 5.4 (c), the detected glottal activity regions cover the burst onset also, due to the presence of high energy. Therefore, it is not reliable to use the detected glottal activity onset points as vowel onsets. Hence, the maximum weighted inner product (MWIP) is used to detect vowel onsets [211]. MWIP evidence is plotted in Figure 5.4 (e). By considering burst onset as a reference, MWIP values are checked for every epoch. The epoch at which the evidence crosses its value above the threshold is referred to as the vowel onset event. The accuracy of the burst detection algorithm is discussed in Subsection 5.2.2.

5.2.2 Performance of burst detection algorithm

The performance of the burst detection algorithm is presented in Table 5.1, which is evaluated by considering a search interval of 30 ms around the manually marked burst location. For the stop /k/, no speech tokens are available corresponding to the palatalized articulation of /k/, and none of the /k/ was uttered correctly. Hence, only the glottal stop substituted /k/ is analyzed. The burst detection performance is evaluated for all the speech tokens reported in Table 3.7 of Chapter 3. The burst

Table 5.1: Performance of burst detection algorithm

Target consonant	Category of error		
	Glottal (%)	Velar (%)	Palatal (%)
/k/	81.98	-	-
/t/	80.12	90.12	90.70
/T/	79.89	91.23	89.87

detection algorithm performs better for the palatalized and velar substitutions than the glottal stop substitution in the case of alveolar /t/ (90.12% and 90.7%) and retroflex /T/ (91.23% and 89.87%). It is because the palatal and velar stops are produced like non-CLP stops, and they contain a strong burst signal. Whereas the glottal stop is produced at the glottis, and soon after the release of vocal folds, the glottal vibrations begin. Hence, the glottal stops may not contain a strong burst component like oral stops. In glottal stops, the sudden onset of voicing is considered a burst-like signal. Here, the burst is produced at the glottis level in the form of creaky phonation. As a result, the glottal stop performance for the three stops (/k/-81.98%, /t/-80.12%, and /T/-79.89%) are observed to be

relatively lower.

Thus, having located the speech signal events, an attempt is made to modify the misarticulated stops in CLP speech. In this study, two ways of event-based speech modifications are considered: burst and burst plus transition. An entire-word modification is also performed using the NMF method for comparative study. The frame wise modification is performed to observe the role of event knowledge in the CLP speech enhancement.

- *Burst modification (B)*: The region to be modified corresponds to a segment between the burst onset point and the VOP.
- *Burst plus transition modification (BT)*: The region to be modified corresponds to a region starting from burst onset point to 20 ms transition region followed by VOP. The main motive of burst plus transition modification rather than a burst-only modification is to modify the acoustic characteristics of consonants embedded in the transition region, which influences the stop consonant's perception [229].

  Automatic stop consonant recognition experiments reported that the modeling of burst plus 20 ms transition region followed by vowel onset results in a better performance. Based on the importance of perceptual cues in the transition region and experimental results of the study [229], a 20 ms region followed by VOP is used for stop consonant modification.
- *Entire word modification (EW)*: The entire speech signal processed based on frame-by-frame transformations without event-knowledge is also studied along with event-based speech modification.

5.3 Spectral transformation based on NMF

In this section, the process of transformation of the stop consonants using the NMF-based spectral transformation method is discussed.

5.3.1 Speech representation

In the NMF-based model, the magnitude spectrum is represented by a linear combination of bases and weights. Each basis is considered an exemplar, and the collection of bases is referred to as a dictionary. The process of estimating dictionaries first involves decomposing the spectrogram, V of dimension F $\times$ K into basis and weights. Assuming I exemplars are collected for an utterance, a dictionary is given by $D = [d_1, d_2, \ldots, d_I] \in R^{F \times I}$, where F denotes feature dimension and d_i

represents the i^{th} exemplar. K denotes the total number of frames in the utterance. Henceforth a speech sample at the l^{th} frame, v_l can be written as,

$$v_l = \sum_{i=1}^{I} d_i h_{il} \approx Dh_l, \quad \text{given } D \geq 0, h_l \geq 0 \tag{5.3}$$

where $h_l = [h_{1l}, h_{2l}, \ldots, h_{Il}] \in R^{I \times 1}$ is the non-negative weight of the i^{th} exemplar. As each speech sample is modeled independently, the spectrogram of an utterance can be approximated in a matrix form as,

$$V \approx DH \tag{5.4}$$

Here, $D \in R^{F \times I}$ and $H \in R^{I \times K}$ are the dictionary and the corresponding weights, respectively. To minimize the distance between V and DH, Kullback-Leibler divergence is used where optimization is acquired based on the multiplicative updating algorithm given in [230].

5.3.2 Spectral transform estimation

Given the source-target spectral pairs, dynamic time warping method is used to aligned the spectral sequences. The source signal corresponds to the CLP speech signals, and the target signal corresponds to the non-CLP speech signals. Prior to the training, both the source and target speech signals are preprocessed and short-time Fourier transform (STFT) is applied on the segmented speech regions. The source magnitude spectrum $x = [x_1, x_2, \ldots, x_K] \in R^F$ is converted into target magnitude spec-

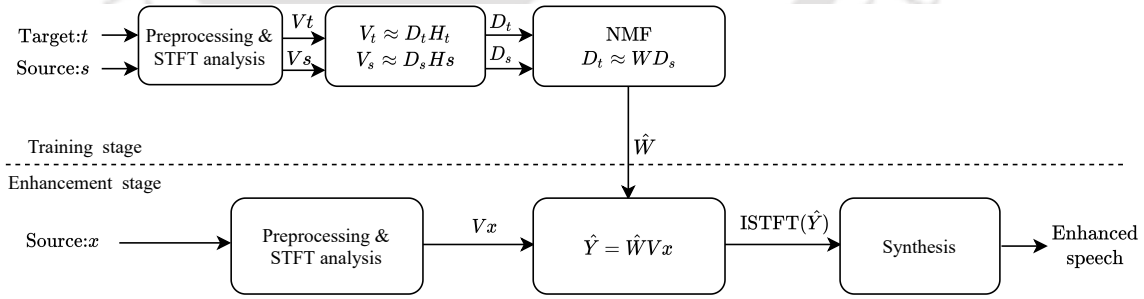

Figure 5.5: Framework illustrating the NMF-based spectral transformation of event-specified misarticulated stops.

trum $y = [y_1, y_2, \ldots, y_K] \in R^F$ using the transformation method given in [231]. NMF is performed at the training stage using parallel source-target spectral pairs. A transformation matrix, W, is learned during the training process. Later at the conversion stage, it is multiplied with the source magnitude spectrum. The transformation matrix is learned from a speaker-independent NMF model. Figure 5.5

shows the process of learning a transformation matrix from the source and target dictionaries, D_s and D_t, which is optimized using the NMF method. D_s and D_t represent the collection of source and target basis, respectively. The source dictionary is constructed using the source features from the disordered spectrogram of CLP speech. The target dictionary is constructed using the target features attained from the spectrogram of non-CLP speech. The two dictionaries consist of aligned magnitude spectral sequences. Given the parallel spectral sequences, $A = D_s$ and $B = D_t$, the target spectral matrix B can be approximated by WA using Kullback-Leibler divergence $\mathscr{D}_{KL}$, given as

$$Z = \mathscr{D}_{KL}(B \,||\, WA) \tag{5.5}$$

The approximation given in equation. 5.5 is minimized by iteratively applying the multiplicative updating rule [230] as follows:

$$W \leftarrow W \otimes \frac{\left(\frac{B}{WA}\right) A^T}{1_{F\times K} A^T + \lambda 1_{F\times F}} \tag{5.6}$$

where W is commonly initialized with an all-ones matrix, $\otimes$ denotes element-wise multiplication, and $1 \in R^{F\times K}$ represents an all-ones matrix. To obtain the converted spectral features ($\hat{Y}$), the source spectral matrix (V_x) is multiplied by the learned transformation matrix ($\hat{W}$), which is given as,

$$\hat{Y} = \hat{W} \times V_x \tag{5.7}$$

An inverse short-time Fourier transform (ISTFT) is applied to the converted spectral sequences and concatenate with the original unprocessed speech segments. Finally, the enhanced signal is synthesized from the converted spectral sequences using the overlap-add method.

5.4 Experimental evaluation

In this section, the misarticulated and the modified stops are evaluated using objective and subjective metrics. Before discussing the details of the subjective and objective evaluations, the impact of the transformation on misarticulated stops is first illustrated in Figure 5.6.

Figure 5.6 depicts the 3-dimensional view of the syllable /ka/ and /ta/ of non-CLP speech, the corresponding CLP speech, and its modified versions. The velar /k/ is characterized by a burst with maximum frequency concentration in the lower frequency region, which can be observed from Figure 5.6 (a) around 10 ms. In contrast, from Figure 5.6 (b) it is observed that CLP /k/ is substituted

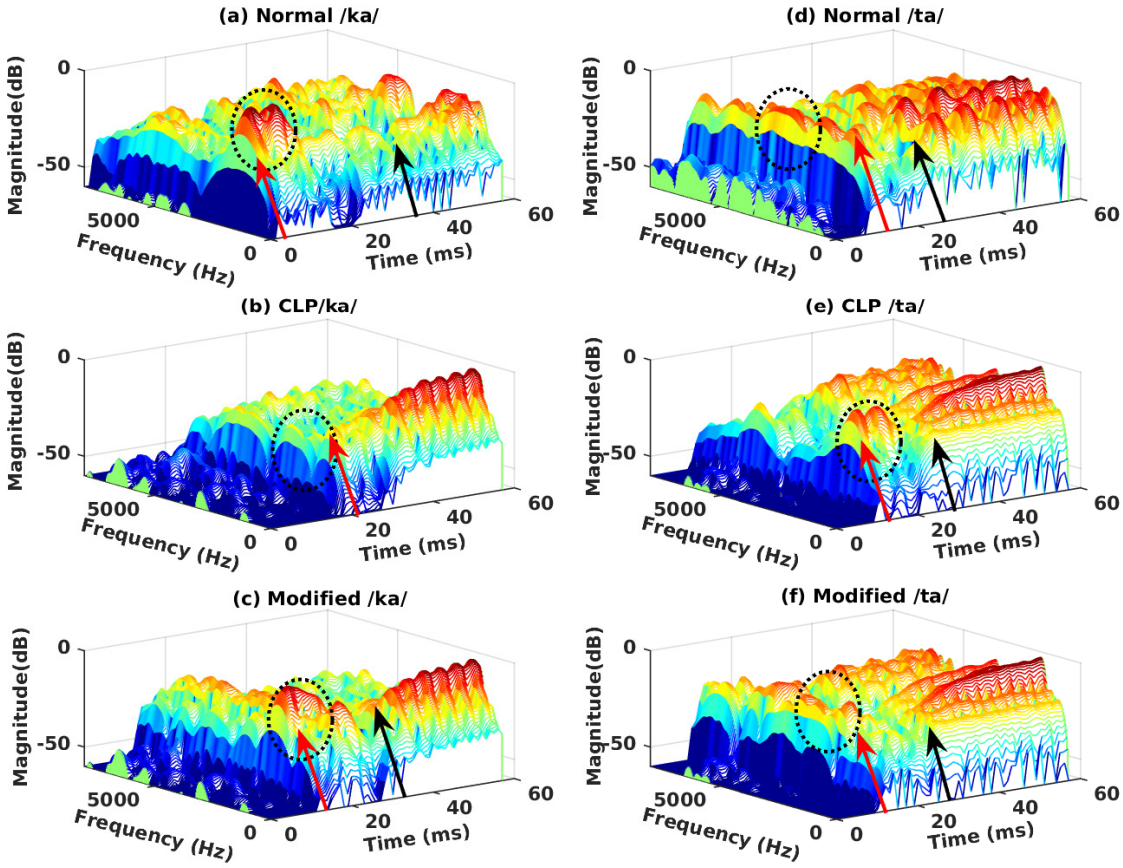

Figure 5.6: Illustration of (a) non-CLP /ka/ syllable with encircled low frequency energy in the /k/ burst, (b) CLP /ka/ syllable where /k/ substituted by glottal stop results in absence of /k/ burst shown by circled region, (c) modified CLP /ka/ syllable, encircled region shows the presence of /k/ burst energy, (d) non-CLP /ta/ syllable, encircled region shows the concentration of maximum energy of /t/ burst in the mid-frequency range, (e) CLP /ta/ syllable, where encircled region shows /t/ substituted by velar /k/ has maximum burst energy in the low-frequency range, and (f) modified /ta/ syllable, encircled region shows the burst energy substitution from low-frequency region to mid-frequency range. Red and black arrows denote the PI and MWIP evidence, respectively.

by a glottal stop where the expected /k/ burst around 15 ms is found to be absent due to the complete adduction of vocal folds. The absence of spectral evidence is observed with an abrupt transition between stop consonant and vowel, and no formant transition occurred between the stop consonant and vowel. The glottal stop is produced at the glottis, and soon after the release of vocal folds, the glottal vibrations begin. Therefore, the glottal stops may not contain a strong burst component like non-CLP stops, as shown in Figure 5.6 (a) and Figure 5.6 (d). In glottal stops, the sudden onset of

voicing is considered a burst-like signal. Hence, in Figure 5.6 (b), only one arrow is shown. Further, from Figure 5.6 (c), it can be observed that the modified stop /k/ shows prominent burst energy (shown by the encircled region) in the lower frequency region. Accordingly, both PI and MWIP evidences are observed after modification.

The velar stop substitution for alveolar stop /t/ and its modified version is illustrated in Figure 5.6 (e) and (f), respectively. From Figure 5.6 (e), it can be observed that velar stop substituted /k/ exhibits spectral prominence in the frequency range of $100 - 2000$ Hz [232]. In the case of non-CLP stop /t/ depicted in Figure 5.6 (d), the burst signal exhibits spectral prominence towards a higher frequency region. After spectral transformation of velar stop substituted /t/ shown in Figure 5.6 (f), the spectral prominence of the burst is shifted towards the high-frequency region. The modified burst is similar to that of non-CLP stop /t/ when the evident spectral energy in the lower frequency region is suppressed.

5.4.1 Objective evaluation

To evaluate the effect of speech transformation, two objective measures are computed: support vector machine (SVM)-based classification and Mahalanobis distance, where the event-based transformed signals are compared with signals that are transformed without using the knowledge of events.

SVM-based classification system

An SVM-based classifier system is developed using a radial basis function (RBF) kernel. The optimum values of the parameters: regularization parameter ($\mathscr{C}$), and RBF kernel width (γ) are experimentally determined using the grid search method. Three separate SVMs are developed for each target (/k/, /t/, /T/), i.e., non-CLP /k/ versus misarticulations produced for the target /k/, non-CLP /t/ versus misarticulations produced for the target /t/, and non-CLP /T/ versus misarticulations produced for the target /T/. Since an SVM is trained for non-CLP and misarticulated stops, the classifier is expected to classify the CLP speech sample before modification as a misarticulated stop. After modification, the speech signal attains non-CLP characteristics, and the modified sample is expected to be classified as a non-CLP stop. The algorithm is considered more successful in modifying if more of the modified stops are classified as non-CLP stops by the SVM classifier.

For each class (/k/, /t/, and /T/), 300 tokens of non-CLP (target) speech are considered as training examples. The misarticulated stop class consists of 228 tokens of /k/, 316 tokens of /t/, and 324

tokens of /T/. Using these samples, three SVMs for the classification of non-CLP and misarticulated stops are trained. The classifier is trained using two-dimensional discrete cosine transform (2D-DCT) features. 2D-DCT features are extracted from a /CV/ transition of 60 ms duration anchored around VOP (i.e., 30 ms on the left and 30 ms right sides of VOP). This 2D-DCT feature-based non-CLP versus misarticulated stop classification approach was proposed in [9]. A 5-fold cross-evaluation is performed in a speaker-independent manner, and the optimum values of $\mathscr{C}$ and γ parameters are estimated. In each fold, four sets are used for training, and the remaining one set is used for testing. In each training phase, the SVM is trained for the different combinations of $\mathscr{C}$ and γ, where, $\mathscr{C} = [2^{-10}, 2^{-8}, \ldots, 2^{+8}, 2^{+10}]$ and $\gamma = [2^{-10}, 2^{-8}, \ldots, 2^{+8}, 2^{+10}]$. At each fold, the parameters for which the model resulted in maximum accuracy are noted. Finally, the $\mathscr{C}$ and γ parameters for which the model yielded the highest accuracy among the 5-folds are noted as the optimum parameters. The averaged accuracy across 5-folds is considered as the accuracy of the model.

The 2D-DCT features computed from the transition regions of original misarticulated, entire-word modification, burst-only, and burst plus transition approaches are given as input to the SVM classifier. The detection rate concerning the non-CLP class is computed and used as an objective measure. Also 70 tokens of /k/, 36 tokens of /t/, and 55 /T/ tokens are used as the non-CLP speech test samples.

The performance of the SVM-based classification system is shown in Table 5.2 for the non-CLP speech and misarticulated CLP speech. Further, the modified speech samples are evaluated using the same system, and the results are also shown in Table 5.2. In Table 5.2, the category of error

Table 5.2: Performance evaluation of non-CLP stops, misarticulated stops, and modified stops using the SVM-based classification system. MS denotes original misarticulated stop, Bm denotes burst modified signal, BTm denotes burst plus transition region modification, EWm denotes entire-word modification, and N denotes non-CLP stop consonants.

Target	Category of error	Target detection rate (%)				
		MS	EWm	Bm	BTm	N
/k/	glottal	14.39	46.67	71.02	83.23	88.89
/t/	glottal	5.34	54.03	8.00	88.58	92.47
	velar	14.52	34.00	14.52	85.00	92.47
	palatal	5.89	24.00	5.89	88.24	92.47
/T/	glottal	7.82	54.33	3.13	71.43	89.81
	velar	25.68	59.74	16.67	71.22	89.81
	palatal	20.00	32.00	51.00	82.50	89.81

is used to denote the type of misarticulation produced for the stops (/k/, /t/, and /T/). From

Table 5.2, it can be observed that the burst plus transition region modified signal (BTm) exhibits a detection rate closer to the non-CLP stop consonants (N) compared to the lower detection rate for the original misarticulated stops (MS). This implies that the transformed speech may have offset the lower performance accuracy for the original CLP speech. From Table 5.2, it can also be observed that the detection rate is relatively lower for burst-only modified signals (Bm) and entire-word modified signals (EWm). Although BTm signals showed a higher detection rate, however, the Bm stop /k/ also shows a relatively higher detection rate compared to /t/ and /T/ consonants. The probable reason for the higher detection rate of stop /k/ may be due to its longer VOT duration compared to stops /t/ and /T/, respectively.

The main purpose of comparing the Bm and BTm with EWm is to observe the role of event knowledge in improving the CLP speech intelligibility. Hence, we compare the speech transformation with and without event knowledge. During the speech transformation, when we use event knowledge, the source spectral components are mapped to the relevant target spectral components. As a result the deviated burst characteristics are modified, which are essential for the stop consonant perception. Whereas, processing the entire speech signal involves block-processing of the signal with fixed window size, where the fixed frame size models might not represent the speech dynamics accurately. As a result, the source spectral components might mix into irrelevant target spectral components, and the quality of the enhanced speech may degrade [136, 146].

Mahalanobis distance

Mahalanobis distance is used as another metric for objective evaluation of the misarticulated and modified stops. Mahalanobis distance gives the signal similarity. It is computed using the equation given in [233].

$$\mathrm{MD} = \sqrt{(x_a - x_b)^{\mathcal{T}} \times C_x^{-1} \times (x_a - x_b)} \tag{5.8}$$

where MD denotes Mahalanobis distance, C_x represents the covariance matrix, x_a represents the vector of the observation obtained from non-CLP speech, x_b represents the vector of the observation obtained from CLP speech, and $\mathcal{T}$ denotes the transpose of the matrix. The distance measures are shown in Table 5.3. It can be seen that compared to original misarticulated stops, the modified stops possess a distance closer to non-CLP stops. The Bm stops and EWm stops have a relatively higher distance closer to original misarticulated stops. This implies that the modified stop consonant acoustic characteristics are similar to that of non-CLP.

Table 5.3: Mahalanobis distance between the target and misarticulated stops and targets and modified stops. MS denotes original misarticulated stop, Bm denotes burst modified signal, BTm denotes burst plus transition region modification, EWm denotes entire-word modification, and N denotes non-CLP speech signal.

Target	Category of error	Mahalanobis distance from the target model				
		MS	EWm	Bm	BTm	N
/k/	glottal	877.61	190.31	211.59	184.91	161.13
/t/	glottal	2108.05	433.44	769.35	321.28	267.24
	velar	1122.15	286.90	429.74	295.12	267.24
	palatal	3232.19	1173.23	422.11	293.02	267.24
/T/	glottal	2241.00	2699.00	605.00	899.00	596.00
	velar	1457.00	714.00	948.00	810.00	596.00
	palatal	1070.00	707.00	650.00	664.00	596.00

5.4.2 Subjective evaluation

For the subjective evaluation of the speech samples, the recruited listeners are research scholars who possess speech technology knowledge. The listeners do not have any hearing problems. Ten listeners are recruited for the evaluation.

Speech quality assessment

To assess the speech quality of the misarticulated stops and the modified stops, a 5-point rating scale mean opinion score (MOS) (1 = bad, 2 = fair, 3 = good, 4 = very good, 5 = excellent) is used. The speech samples were randomly numbered to avoid any bias towards the original or modified speech. In the listening test, /k/ misarticulation and three types of /t/ and /T/ misarticulations modified using the NMF method are provided to the volunteer listeners. All the listening is made through headphones. For each of the misarticulations, ten modified speech utterances and ten original speech utterances are presented. A detailed description of the misarticulation and the expected target word is explained to the volunteers before performing the test to avoid the listener's distraction due to the nasalization of vowels and abnormalities in pitch, loudness, and voice quality. If they consider other distortions in the /CVCV/ word, then they might not evaluate the target appropriateness, instead, they will evaluate entire-word-level intelligibility, which is not addressed in this work. The MOS scores are evaluated by SLPs and naive listeners. The mean and standard deviation MOS scores are shown in Table 5.4 and Table 5.5, respectively. Compared to Bm and EWm, the tables show that BTm

Table 5.4: Mean opinion scores of the original and modified stop consonants evaluated by naive listeners. MS denotes a misarticulated stop, and Bm denotes a burst modified signal. BTm denotes burst plus transition region modification, and EWm denotes entire-word modification.

Target	Category of error	Mean opinion score (MOS)			
		MS	EWm	Bm	BTm
		$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$
/k/	glottal	1.18±1.05	1.30±0.85	3.16±0.31	3.25± 0.36
/t/	glottal	1.13±1.11	1.43±1.12	2.33± 0.99	3.50±0.77
	velar	1.07±1.21	1.60±1.04	2.40± 0.87	2.82±0.92
	palatal	1.28±0.79	2.87±1.16	2.42±0.97	3.57± 0.99
/T/	glottal	1.17±1.33	1.14±1.31	2.14±0.81	2.95±0.82
	velar	1.10±1.03	2.30±1.11	2.00±1.26	3.13±0.83
	palatal	1.30±0.89	2.20±0.75	3.00±0.57	3.60±0.59

Table 5.5: Mean opinion scores of the original and modified stop consonants evaluated by the speech-language therapists (lay listeners). MS denotes original misarticulated stop, and Bm denotes burst modified signal, BTm denotes burst plus transition region modification, and EWm denotes entire-word modification.

Target	Category of error	Mean opinion score (MOS)			
		MS	EWm	Bm	BTm
		$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$	$(\mu \pm \sigma)$
/k/	glottal	1.02±0.99	1.50±1.01	3.50±1.00	3.50±0.98
/t/	glottal	1.30±1.12	1.20±0.97	2.20±0.85	3.65±0.81
	velar	1.00±0.87	1.30 ±0.58	2.70±0.36	2.98±0.32
	palatal	1.00±0.90	1.70±0.98	2.60±0.85	3.30±0.87
/T/	glottal	1.20±1.11	1.40±1.02	2.01±0.97	3.10±0.54
	velar	1.20±1.05	1.90±1.39	2.50±0.95	3.50±0.83
	palatal	1.03±0.93	1.50±0.98	2.90±0.88	3.40±0.81

signal exhibits higher scores. The MOS score of approximately 3 for BTm signals indicates that the perceptual characteristics are transformed, which tend to be like non-CLP speech. Comparatively, the original CLP speech possesses a MOS score around 1. The MOS of approximately 3 for the BTm signal implies that the speech quality can be further improved. The reason for lower MOS values may be attributed to the distortions of the vowels in the /CVCV/ word and other transformation parameters involved in modifying the speech segments.

Preference test on similarity

The modified signal performance is also evaluated in terms of subjective perceptibility on the modified speech signal similarity with the source and target speech signals. In this test, a 4-scale score (1 =

different and absolutely sure, 2 = different and sure, 3 =same but not sure, 4 = same and absolutely sure) [231] is used to evaluate the similarity of the modified speech with the misarticulated signal and non-CLP signal. Here, ten listeners listened to 3 sets each: Set A, Set B, and Set C, respectively. In Set A, two similarity tests are performed, one corresponds to the similarity test between the EWm & MS, and another corresponds to the similarity test between EWm & N signals. Five pairs of EWm & MS and five pairs of EWm & N corresponding to each of the stop consonants for each category of error are provided to the listeners. Thus, in each set, a total of 70 ($5 \times 7 + 5 \times 7 = 70$) pairs of speech signals are evaluated. Set B is also used to evaluate two similarity tests, where one corresponds to the similarity between Bm & MS, and another one corresponds to Bm & N. Here also, five pairs of Bm & MS signals and five pairs of Bm & N signals corresponding to each of the stop consonant and category of error are provided to the listeners. Accordingly, Set C is also used to evaluate two similarity tests, between BTm & MS signals, and between BTm & N signals. Therefore, in Set C, five pairs of BTm & MS and five pairs of BTm & N corresponding to each of the stop consonant and category of error are provided to the listeners. In each pair, the order of utterances is randomized.

Table 5.6: Preference tests on the similarity of the modified stop consonants with target stop consonants and misarticulated stop consonants. MS denotes original misarticulated stop, Bm denotes burst modified signal, BTm denotes burst plus transition region modification, EWm denotes entire-word modification, and N denotes non-CLP speech signal.

Target	Category of error	Similarity Score					
		Set A		Set B		Set C	
		EWm-MS	EWm-N	Bm-MS	Bm-N	BTm-MS	BTm-N
/k/	glottal	2.35	1.23	1.94	2.60	1.05	3.45
/t/	glottal	2.25	1.31	2.00	2.16	1.69	2.56
	velar	2.54	1.76	2.52	2.89	1.42	3.26
	palatal	3.71	2.39	2.43	3.02	1.39	3.18
/T/	glottal	3.71	1.21	3.20	2.39	2.12	2.78
	velar	2.10	2.19	2.78	1.60	2.30	3.82
	palatal	3.88	2.00	2.62	3.00	2.10	3.09

The evaluators were asked to give a preference based on the 4-scale score for the modified speech signal, compared to the target signal and source signal. Suppose the evaluators do not perceive any similarity neither with the source nor with the target signal. In that case, they can mark it as "No preference". The results of preference assessment on similarity are shown in Table 5.6. It can be observed from all the three sets that Set C exhibits the lowest similarity scores for modified and misarticulated signals and relatively higher similarity scores for the modified and non-CLP signals.

From Set B and Set A, it is speculated that the Bm signal and an EWm signal show close similarity with the source signal.

5.5 Summary

This work proposed an event-based spectral transformation of articulatory errors for the intelligibility improvement of CLP speech [207]. Prior to modification, the burst and vowel onset events are located by first detecting the glottal activity regions. Having located the events, transformation models are developed for the three-stop consonants: /k/, /t/, and /T/, which are used to transform the misarticulated stops. NMF-based spectral transformation is used to modify the stop consonants in CLP speech. The SVM classifier results revealed that a BTm signal has a higher detection rate than a Bm signal and an EWm signal. Like SVM classifier results, Mahalanobis distance also showed the same trend of distance measures, where BTm signal shows the lowest distance compared to original misarticulated and other modified signals. The subjective MOS rating of BTm signal is approximately "good", and it implies that the modified speech signal has lower spectral distortion. Furthermore, the SVM classifier results and Mahalanobis distance reveal a similarity between the modified speech and non-CLP speech. The BTm signal has a lower Mahalanobis distance for a higher detection rate. A higher similarity score implies that the modified signal has perceptually similar acoustic characteristics to the non-CLP signal.

For a /CVCV/ word enhancement, apart from consonants, the vowel distortion also needs to be addressed because nasalized vowels (hypernasality) are observed to influence the speech perceptivity. Hence, vowel modification is performed in the following chapter.

6

Modification of hypernasal vowels using temporal and spectral processing

Publications

[234] Protima Nomo Sudro, S. R. Mahadeva Prasanna,"Enhancement of cleft palate speech using temporal and spectral processing", *Speech Communication* 123 (2020): 70-82.

[208] Protima Nomo Sudro, " Intelligibility Enhancement of Cleft Lip and Palate Speech", 5^{th} *Doctoral consortium, Interspeech* 2019.

Contents

Overview

The speech of the individuals with cleft lip and palate (CLP) is generally characterized by the presence of abnormal nasal resonances during the production of voiced sounds, primarily in vowels. This phenomena is called hypernasality. Hypernasality is present in more than 50% *of the individuals with CLP. It often results in degraded speech, with poor quality and intelligibility. This work describes the signal processing based enhancement of CLP speech, where specifically hypernasal speech modification is addressed. The hypernasal speech is parameterized using an extended weighted linear prediction (XLP) method. The enhancement is performed for three different variants: XLP residual weighting in the time domain, Gaussian mixture model-based spectral conversion in the frequency domain, and combined modification of the XLP residual and vocal tract system characteristics. The modified hypernasal speech achieved by the proposed method is evaluated using different objective and subjective measures for the vowel /a/, /i/, and /u/, respectively. The evaluation results indicate that the combination of XLP residual and vocal tract system characteristics modification yields better results than XLP residual or vocal tract system characteristics modification alone.*

6.1 Introduction

Hypernasality corresponds to a resonance disorder, and it is a primary disorder observed in CLP speech. The presence of nasal resonances during the production of oral sounds results in a deviant speech that has an excessively perceptible nasal quality [4]. The introduction of nasal resonances is caused by the presence of velopharyngeal dysfunction (VPD) and/or oronasal fistula. The nasalization of vowels reduces the clarity of speech, thus distorting both the quality and intelligibility [2, 235]. A study reported that intelligibility decreases with an increase in hypernasality and suggested that hypernasality interacts with intelligibility [12]. The distorted speech patterns caused by hypernasality make the individuals with CLP less confident when engaging in conversational skills compared to the speakers with typical speech and language [13, 15, 236]. The clinical treatment for correcting hypernasal speech involves primary and secondary surgery. Although surgical management should result in the elimination of hypernasality, distorted speech patterns persist even after surgery in most of the individuals [237].

From the studies of hypernasal speech in the literature, it is observed that explicitly, vowel characteristics are deviated due to the introduction of additional formant and anti-formant pairs in the

spectrum, broadening of formant bandwidths and spectral flattening [238,239]. The hypernasal speech characteristics are also studied and analyzed based on the multicomponent nature using the Teager energy operator and the combination of the Teager energy operator plus Mel frequency cepstral coefficient (MFCC) feature. [240,241]. The distance between low and high order linear prediction cepstral coefficient is used as a parameter for the detection of hypernasality [242]. To resolve the spacing of the nasal formant in the vicinity of the first formant in the hypernasal vowel spectrum, an acoustic measure based on the group delay spectrum is proposed in [243]. Additionally, zero time windowing is also used for the analysis of hypernasal speech [244,245]. Hypernasality is also characterized using features, namely, acoustic, noise, cepstral analysis, non-linear dynamics and a combination of non-linear dynamic features plus entropy measurements [246–248]. The characteristics of hypernasal speech are also analyzed using the Rademacher complex model for two Spanish words /koko/ and /papa/ [249]. Further hypernasal speech detection is performed in a sentence speech using jitter, shimmer, MFCC bionic wavelet transform energy, and bionic wavelet transform energy [250]. Studies showed that hypernasality detection and severity analysis is performed using the vowel space area [251, 252]. The dominance of low-frequency energy in hypernasal vowel /a/ is investigated using voice low tone to high tone ratio (VLHR) [253]. The energy distribution concept is explored for three vowels /a/,/i/ and /u/ in [254].

6.1.1 Challenges

Motivated by the importance of hypernasal speech enhancement, the aim of this chapter is to improve vowel perceptibility in the hypernasal speech by reducing the distortions caused by nasalization. Generally, hypernasality is assessed by using a test word composed of /CV/ structure [8]. Therefore, the words in the form of /CV/ structures are considered for hypernasal speech enhancement, where the V unit corresponds to elongated vowel phonation. The emphasis is given on vowel modification because hypernasality is a resonance disorder, and it is particularly perceived in voiced sounds, especially vowels. The vowel sounds are produced by altering oral resonances associated with higher energy and are relatively long in duration. The details of the hypernasal speech data used in this work are presented in detail in Subsection 3.3.3 of Chapter 3.

The studies reported above revealed that the spectral characteristics of the vocal tract system are mostly deviated, which leads to speech quality degradation. It is also reported that, because of the deviated spectral characteristics, the obtained residual signal consists of scaled and delayed

versions (interference/additional components other than significant glottal closure instants) of the original speech [255]. During synthesis, the residual signal will be used to excite the time-varying all-pole filter. While synthesis, the interference/additional signal components in the residual signal may sometimes introduce unnatural spectral changes, which are perceived as distortion in the enhanced speech. Therefore, the interfering distortions in the residual signal must be minimized to avoid inaccurate formant estimation, which causes significant distortion in the resulting enhanced speech. Thus, modification of residual and spectral characteristics are expected to result in enhanced speech.

Although conventional linear prediction (LP) analysis exhibits several benefits for extracting a smooth spectral envelope. However, it is noted to be inaccurate while estimating the formant frequencies of high-pitch speech. The degradation is caused by the least square error criterion used in the LP method [256]. The spectral envelope of high-pitch speech, extracted using cepstral analysis, is also observed to be affected by the pitch harmonics [257, 258]. It is reported to dampen the low-frequency region, which is important for hypernasal vowel analysis. The speech transformation and representation using adaptive interpolation of weighted spectrum (STRAIGHT) algorithm is also extensively used to extract smooth spectral envelope in the voice conversion (VC) area. The STRAIGHT method is widely used for parameterization because of its properties in isolating the $F0$ effects from the vocal tract. However, in a study, it is also reported that the higher $F0$ value and higher formant frequencies of children speech might affect the accurate estimation of the spectral envelope [172].

6.1.2 Contributions

In the case of high-pitch hypernasal speech, more accurate spectral envelope estimation is important because the effect of $F0$ may reduce the discriminability between nasal formant and the harmonics. Additionally, nasal formant occurs in the oral formant vicinity, and it reduces the discriminability between nasal and oral formants. Therefore, the extended weighted linear prediction (XLP) method can be used to avoid the biasing effect of $F0$ in high-pitch speech. Compared to the conventional approaches (for example, LP method, cepstral analysis, and STRAIGHT) of extracting smooth spectral envelope, XLP analysis allows more versatility in emphasizing the speech components during the computation of optimal filter coefficients.

Most of the speech enhancement approaches mentioned in Chapter 2 were proposed for noisy speech modification with an assumption that noise is additive and uncorrelated with a clean signal. In order to use the speech enhancement methods proposed for noisy speech modification, we need to

estimate the following attributes:

- Nasal spectrum from the high pitched speech and remove it from the hypernasal spectrum to obtain a non-nasal spectrum.
- Estimate the nasal and non-nasal amplitude.
- Decompose the vector space of hypernasal speech into non-nasal and nasal subspace.
- Estimate other nasal and non-nasal speech parameters.

However, unlike noisy speech, in hypernasal speech, nasality is not a distinct component of speech, but it is produced from the same vocal tract system and correlated with speech. As a result, coupling occurs between the nasal cavity and oral cavity, due to which additional formant and anti formant pairs are observed in the lower frequency region of the spectrum.

Henceforth, a method is proposed for hypernasal speech modification. Here, the spectral transformation consists of mapping the extended weighted linear prediction cepstral coefficients (XLPCC) of hypernasal speech towards the target XLPCC by using a trained conversion function. The modified XLP residual signal is used to excite the time-varying all-pole filter derived from transformed spectral characteristics of hypernasal speech to produce enhanced speech.

The rest of the chapter is organized as follows: Section 6.2 provides an overview of the transformation method and analysis of hypernasal speech using XLP method. In Section 6.3, the modifications of XLP residual and XLPCC are dicussed. Experimental results and discussions representing the objective and subjective evaluations are described in Section 6.4. Section 6.5 describes the summary of the work.

6.2 Methodology

In this work, the proposed approach is carried out by first analyzing the hypernasal speech followed by transforming the XLP residual and vocal tract system characteristics. The details of the metadata used in this work is described in Subsection 3.3.3 of Chapter 3. Finally, the speech is synthesized using modified speech features.

6.2.1 XLP based analysis of hypernasal speech

XLP is a generalized formulation of the LP method with temporally weighting methods as special cases [259,260]. From Chapter 1 and aforementioned literatures on hypernasal speech, it is noted that more severe speech deviations are observed based on the degree of cleft severity. In case of severe

deviations, the speech mostly consists of nasal sounds such as /m/, /n/, and /ng/ [261, 262]. Each of the spectral components of the vowels has a different influence on perceived hypernasality [74]. Therefore, an accurate representation of the spectral envelope is necessary due to following attributes:

- In high-pitch speakers, lowest formants are biased by pitch harmonics.
- Hypernasal speech introduces additional resonances at various frequency locations with consistent nasal resonances in the low-frequency region.

The biased formant estimation problem for high-pitch vowels can be overcome using XLP method [263]. It temporally weights the speech samples [259, 260]. By weighting separately on each lagged signal sample in the prediction model, it has shown better performance for cases with channel distortion mismatch and low to moderate additive noise. XLP method provides a robust spectral estimation technique, and its model stability is guaranteed when the weights are defined recursively.

Extended weighted linear prediction method

The speech samples are weighted in the prediction model, and the prediction coefficients are optimized according to the least square error criterion. The estimate of XLP parameters are solved by minimizing the prediction error energy E_{XLP} as,

$$E_{XLP} = \sum_{n}(s_n Y_{n,0} - \sum_{k=1}^{p} d_k s_{n-k} Y_{n,k})^2 \tag{6.1}$$

where, s_n, s_{n-k}, d_k and $Y_{n,k}$ denote the current speech sample, previous speech sample, predictor coefficients and the weighting function, respectively. n denotes the sample number and k denotes the predictor coefficient index. The weighting function is used to emphasize the spectral regions of prominent energy. The XLP model is obtained by solving the normal equations given by,

$$\sum_{k=1}^{p} d_k \sum_{n} Y_{n-k} s_{n-k} Y_{n,j} s_{n-j} = \sum_{n} Y_{n,0} s_n Y_{n,j} s_{n-j}, 1 \leq j \leq p \tag{6.2}$$

In order to compute the weights, the following recursion is used

$$Y_{n,j} = \frac{m-1}{m} Y_{n-1,j} + \frac{1}{m}(|s_n| + |s_{n-j}|) \tag{6.3}$$

The parameter m controls the effective length of the moving average memory. Here, the length of m is equivalent to the linear prediction order, which is 14 in this work. The weighting function $Y_{n,j}$ is specified as the absolute value sum. The underlying rationale for the weighting function is that it will

emphasize the less distorted higher amplitude samples more compared to lower amplitude samples. The optimal d_k values from equation 6.2 is used to obtain the inverse filter of the XLP analysis,

$$A(z) = 1 - \sum_{k=1}^{p} d_k z^{-k} \tag{6.4}$$

Model stability is guaranteed when $Y_{n,j}$ are defined recursively as,

$$Y'_{n,j} = max(Y_{n,j}, Y_{n-1,j-1}) \tag{6.5}$$

with $Y_{n,j} = 0$ for $j < 0$. The resulting analysis method is denoted as a stabilized XLP method. From equation 6.1 and 6.2, separate temporal weighting are observed for the present and past speech samples, respectively. The spectral representation is approximated to be free from the variations of fundamental frequencies because of the separate temporal weighting of the prominent speech samples. Efficacy of the spectral envelope estimation using the XLP method compared to conventional LP and fast Fourier transform (FFT) method is studied.

Effectiveness of XLP method in the spectral analysis of high-pitch speech

The magnitude spectra estimated from the high-pitch speech of non-CLP (healthy) and hypernasal speakers are demonstrated for the vowel /i/ in Figure 6.1 (a)-(d) for non-CLP speech and in Figure 6.1 (e)-(h) for hypernasal speech. It is observed that the resonances can be identified from LP, XLP, and FFT spectra as prominent local peaks from both non-CLP and hypernasal speech. However, the resonances estimated by LP methods either move up or down in frequency with an increase in $F0$. In the case of the XLP method, a consistent trend is observed where the formant frequency increases when $F0$ increases. These varying locations of the resonances may be attributed to the fact that the LP method is sensitive to the biasing effect of the $F0$. Whereas the corresponding resonances computed by the XLP method are consistent with F0 variation. In Figure 6.1 (e)-(h), additional nasal resonances are also observed around 1000 Hz as compared to the magnitude spectra of non-CLP speech in Figure 6.1 (a)-(d). Since the XLP method estimates the formant frequency locations more accurately than the conventional LP method. Therefore, in this work, the XLP method is used to estimate residual and vocal tract system characteristics of hypernasal speech.

Analysis of XLP residual

The XLP residual signal is estimated by filtering the hypernasal speech signal through the filter $A(z)$ of

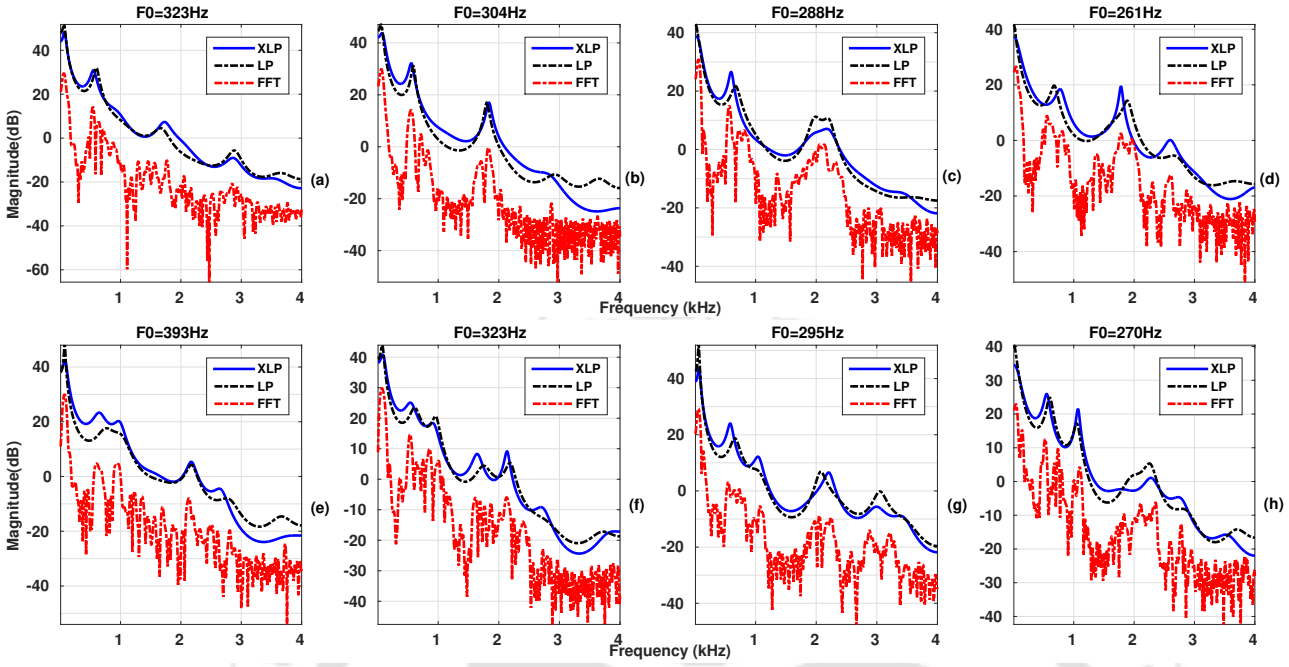

Figure 6.1: Illustration of magnitude spectra for high-pitch speech computed using conventional LP, XLP, and FFT for vowel /i/ phonation of varying F0 of (a)-(d) non-CLP speakers and (e)-(h) hypernasal speakers.

equation 6.4. The XLP residual signal for vowel /i/ phonation of non-CLP and hypernasal speech is

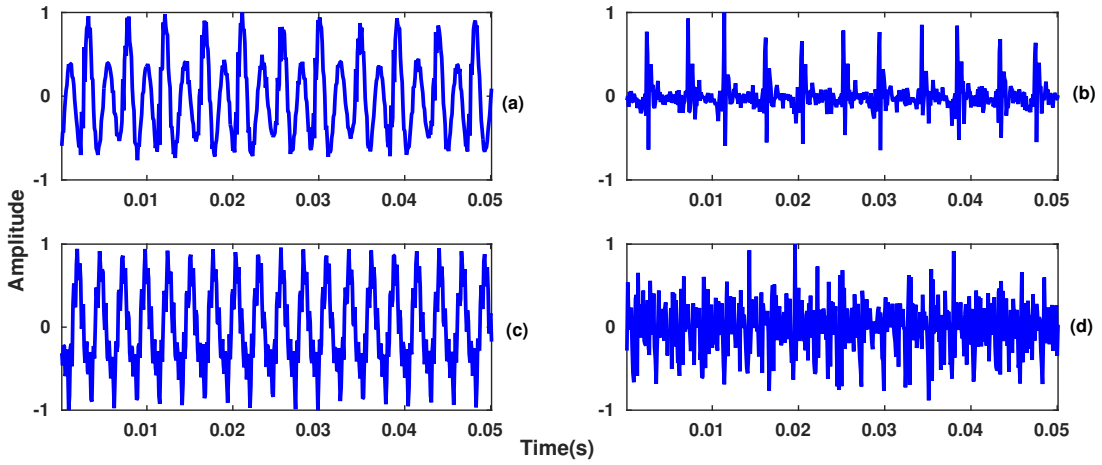

Figure 6.2: A segment of vowel /i/ phonation of (a) non-CLP speaker and corresponding (b) XLP residual, (c) hypernasal speaker and corresponding (d) XLP residual.

shown in Figure 6.2. From Figure 6.2 (d), it can be seen that significant interfering signal components arise around the glottal closure instants (GCIs) in the XLP residual of hypernasal speech. However, in Figure 6.2 (b), such significant pulses are comparatively very low. The impulse-like excitations

during the closing phase of a glottal cycle are represented as GCIs [264]. The corresponding strengths of GCIs are significantly larger relative to other regions of the signal. To quantitatively analyze the significant interfering signal components concerning the significant GCI locations, the peak-to-sidelobe ratio (PSR) [265] of the Hilbert envelope of the XLP residual is analyzed. The peak represents the value of the central peak at the GCI location, and sidelobes represent a 1.5 ms duration segment towards the right of the GCI, respectively. Because of the interference XLP residual, the GCIs are not always prominent. Therefore, a robust method is required to locate the instants of significant excitations from the hypernasal speech signal. It is reported in the literature that, zero frequency resonator, generally termed as zero frequency filter (ZFF) gives reliable estimates of GCIs. Hence, ZFF is used to estimate GCIs [197]. The filtered signal is known as the ZFF signal, and it exhibits discontinuities at GCI locations at positive zero crossings. After locating the GCI locations, the PSR

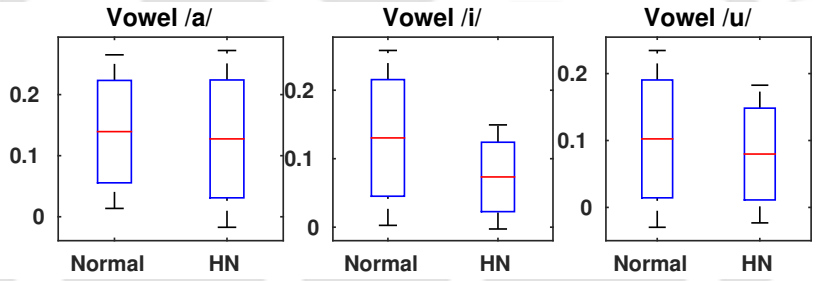

Figure 6.3: Boxplot showing peak-to-sidelobe ratio (PSR) values for the three vowels of non-CLP and hypernasal speakers. HN denote hypernasality.

is now computed for hypernasal and non-CLP speech and it is depicted in Figure 6.3. From Figure 6.3, it can be noted that the PSR values of the hypernasal vowels /i/ and /u/ are significantly lower than the non-CLP vowels. Whereas, the difference between non-CLP and hypernasal vowel /a/ is less significant because vowel /a/ is a low vowel, and it is less affected due to the nasalization compared to high vowels. It is observed that non-CLP vowels have higher PSR value compared to the hypernasal vowels. One possible explanation for this behavior is the presence of interfering signal components in the XLP residual of hypernasal speech. The introduction of interfering signal components in the XLP residual corresponds to the presence of nasal zeros caused by hypernasality. The introduction of nasal zeros in the transfer function affects the accuracy of the estimation of formants. An inaccurate formant estimation results in prediction error coefficients. The additional prediction error coefficients, introduce scaled and delayed versions of the original signal in the XLP residual [255]. Thus, a higher side-lobe values are noted for the hypernasal speech resulting in lower PSR values.

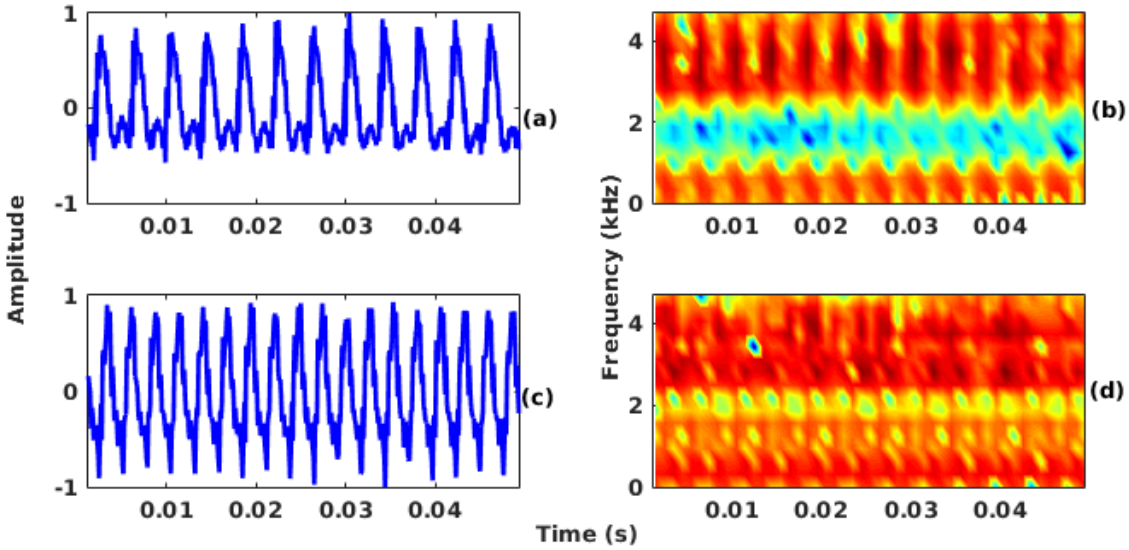

Figure 6.4: Comparison of the speech signal and corresponding spectrogram of (a)-(b) non-CLP speaker and (c)-(d) hypernasal speaker for vowel /i/ phonation.

Analysis of vocal tract system characteristics

As hypernasality primarily corresponds to a resonance disorder, analysis of vocal tract system characteristics is carried out to characterize the spectral deviations. The spectral analysis reflected the deviation regarding additional nasal resonances along with the oral cavity resonances. In Figure 6.4 (a)-(d), non-CLP and hypernasal speech signal and the corresponding spectrograms are illustrated for the vowel /i/ phonation. Comparable differences are observed in the spectral energy levels of Figure 6.4 (b) and Figure 6.4 (d), respectively. The prominent spectral energy represents the resonances of the signal. From the hypernasal spectrogram depicted in Figure 6.4 (d), it is observed that the energy level of first resonance and between the first and second resonance are relatively higher than those in Figure 6.4 (b). As a result, the clarity/ quality of the vowel is substantially reduced. The spectral changes of a hypernasal vowel observed in Figure 6.4 are further analyzed by estimating VLHR.

It is an objective index used to detect the nasalization effect from the spectral characteristics of voiced signals, particularly vowels [253]. The voice spectra are divided into a low-frequency band (LFB) and a high-frequency band (HFB) with a cut-off frequency, $F_c = 600$ Hz. The LFB and HFB are defined as the summation of power of each of the frequency component, ranging from 65 to F_c Hz and F_c to $\frac{f_s}{2}$ Hz respectively, f_s is the sampling frequency, which is equal to 10 kHz. Mathematically, VLHR is expressed as, $VLHR = 10 \times log_{10}(\frac{LFB}{HFB})$. To estimate overall VLHR, the VLHR values are

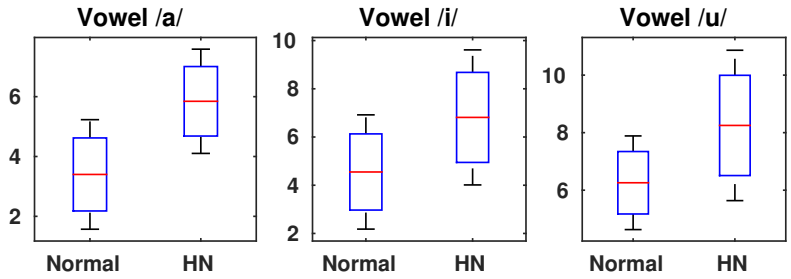

Figure 6.5: Boxplot showing voice low tone to high tone ratio (VLHR) for the phoneme /a/, /i/ and /u/ of non-CLP and hypernasal speakers. HN denotes hypernasality.

averaged across all speech frames for each vowel. From Figure 6.5, it is noted that for all the three vowels, there is clear discrimination between non-CLP and hypernasal speech. The presence of nasal resonances in the low-frequency region results in a significant increase in VLHR values for all the three vowels of hypernasal speakers. Since both the XLP residual and vocal tract system characteristics are essential for speech quality and intelligibility, modification of both aspects is performed to achieve enhanced speech.

6.3 Hypernasal speech enhancement

The block diagram of the proposed hypernasal speech enhancement method is illustrated in Figure 6.6. All the speech signals are analyzed by short-time processing preceded by windowing a frame of 25 ms duration with 5 ms overlapping interval. Before the short-time processing, speech signals are down-sampled at 10 kHz. During the training process, an analysis is performed on the source

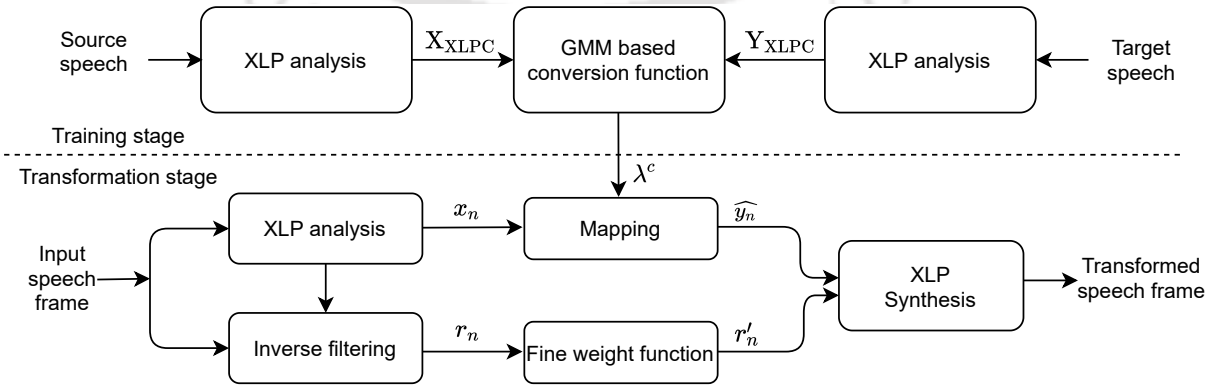

Figure 6.6: Illustration of the framework of the vowel enhancement method. XLP is the extended weighted linear prediction, XLPCC denotes extended weighted linear prediction coefficient cepstrum, and XLPR is the extended weighted linear prediction residual.

and target speaker's utterances to derive the feature parameters to be transformed. In this work, XLPCCs are used as the feature parameter, which is then used to build the conversion function. A 14^{th} order XLPC and XLPCC are used in the study. The conversion function is built for the vocal tract transfer function. During the transformation process, similar feature parameters that were extracted in the training stage are derived from the incoming speech samples. The feature parameters are then transformed based on the conversion function from the training stage. Further, the additional interfering components in the extended weighted linear prediction residual (XLPR) is suppressed by convolving a fine weight function around the significant GCIs. The continuous waveforms are obtained by concatenating the short-time modified speech using the overlap and add method.

The speech data is prepared into several rotations of training and testing sets. The training set consists of all the frames of the corresponding vowel phonations from non-CLP and hypernasal speakers. Leave-one speaker out cross-validation is performed to avoid any biases. In this work, 39 severe hypernasal speakers speech is analyzed and performed enhancement accordingly. The assignments are performed using random permutation. Based on the random number generator with 39 severe hypernasal speakers, 39 different rotations of training and testing sets are obtained.

Hypernasal speech enhancement is carried out for three variants, namely suppression of the interfering components of XLP residual $l(n)$, GMM based spectral conversion $v(n)$ of the vocal tract system characteristics and enhancement based on the combination $s_{tf}(n)$ of the first two variants. The different variants of enhancement are studied to analyze the influence of hypernasality on the XLP residual and vocal tract system components. Therefore, to modify the XLP residual signal and vocal tract system characteristics, the speech signal is resolved in terms of source and filter components using XLP analysis, which is presented in Section 6.2.1.

6.3.1 Temporal processing of hypernasal speech

The XLP residual signal is uncorrelated and any modification performed will result in the least distortion for the synthesis of the speech signal. Hence, the rationale behind the XLP residual modification is to emphasize the significant excitations represented by GCIs and suppress the interference/additional components from the residual signal samples. To suppress the interfering signal components of XLP residual, the instants of significant excitation extracted using ZFF are used as anchor points. A fine weight function is derived similar to Ref. [266] for modifying the XLP residual. Then, the region around GCI location is convolved with a hamming window function, h_{w_n} of 3 ms duration.

By considering GCIs as the shifted train of impulses, the fine weight function, w_{f_n} is expressed as,

$$w_{f_n} = \sum_{k=1}^{T_k} \delta_{n-i_k} * h_{w_n} \tag{6.6}$$

where T_k represents the total number of GCIs, and i_k denotes the GCI index. To avoid over-emphasizing GCI locations in XLP residual, the minimum value of w_{f_n} is set to a threshold value of T_r, and it is set to 0.5. It is reported that temporal processing is not sensitive to T_r values [266]. Therefore, the relationship is expressed as,

$$w_{f_n} = \begin{cases} T_r, & W_{f_n} < T_r \\ w_{f_n}, & otherwise \end{cases} \tag{6.7}$$

The weighted XLP residual, r_{wn} is obtained by multiplying the fine weight function, w_{f_n} with the

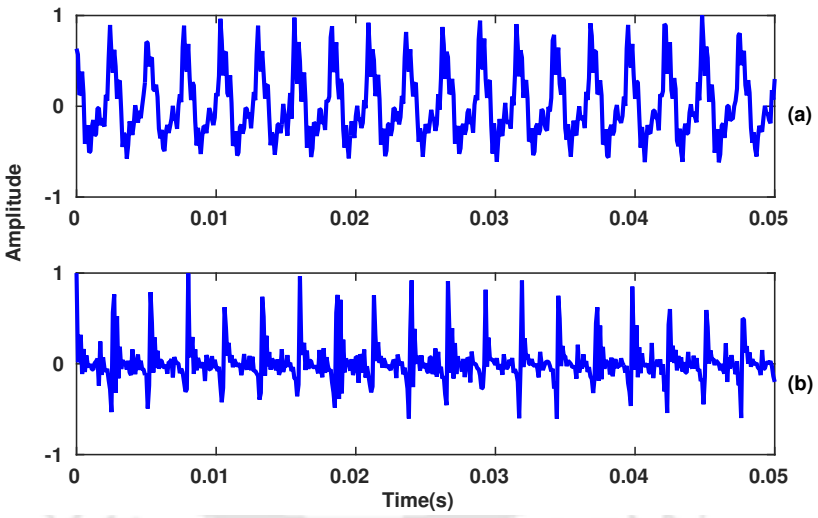

Figure 6.7: Illustration of (a) modified vowel /i/ phonation and the corresponding (b) XLP residual.

hypernasal speech residual signal which is represented by prediction error, r_n as

$$r'_n = w_{f_n} \otimes r_n \tag{6.8}$$

where $\otimes$ denotes element-wise multiplication. The weighted residual is used to excite the all-pole filter derived from hypernasal speech to generate the temporally processed speech signal using the transfer function in Z domain as,

$$L(z) = \frac{R'(z)}{1 + \sum_{k=1}^{p} d_k z^{-k}} \tag{6.9}$$

where, $L(z)$ is obtained using modified XLP residual $R'(z)$, which is weighted XLP residual and d_k denote the predictor filter coefficients of the hypernasal speech. The temporally enhanced speech signal and the enhanced XLP residual signal is shown in Figure 6.7 (a) and Figure 6.7 (b) respectively. Comparing Figure 6.7 (b) with Figure 6.2 (d), it is observed that the significant interfering signal components other than the impulse-like excitations are suppressed. However, after the reduction of significant interfering components from the XLP residual, the additional resonances of the vocal tract system persists. This implies that during reconstruction, the all-pole filters derived from the hypernasal speech dominates the vocal tract characteristics in the enhanced speech signal. It necessitates the need to improve the vocal tract characteristics at the spectral level.

6.3.2 Spectral processing of hypernasal speech

The modification of the XLP residual described in Section 6.3.1 is observed to deemphasize the additional interfering components from the XLP residual signal samples. However, considering the fact that hypernasality is intact in the vocal tract resonances, the characteristics of resonance structure must also be modified to obtain enhanced speech.

GMM based spectral conversion

For the spectral conversion, first, the source speaker and target speaker XLPC cepstra are derived. Then, a GMM model is fitted to the augmented source speaker and target speaker features. The joint probability density of the source and target XLPC cepstrum is expressed as,

$$P(\mathrm{c}_n|\lambda^{\mathrm{c}}) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathrm{c}_n; \mu_m^{\mathrm{c}}, \Sigma_m^{\mathrm{c}}) \tag{6.10}$$

where, w_m denotes the prior probability that the vector belongs to the m^{th} class, $\sum_{m=1}^{M} w_m = 1,\ \ w_m > 0$. $\mathcal{N}(\mathrm{c}_n; \mu_m^{\mathrm{c}}, \Sigma_m^{\mathrm{c}})$, denotes the Gaussian distribution with mean vector μ and covariance matrix Σ, m is the mixture component index, M is the total number of mixture components and λ^{c} is a parameter set of the GMM trained on joint vector c_n. The parameters are initialized by the use of binary splitting vector quantization (VQ) procedure. The joint vector, $\mathrm{c}_n = [x_n^{\mathcal{T}}, y_n^{\mathcal{T}}]$ consists of vector sequences of x and y. $\mathcal{T}$ represents the transposition of the XLPC cepstrum vector. The vector $\mathrm{c}_n(n = 1, 2, \ldots, N)$ represent the time-align sequences with N as the total frame number of the training data for the given speech corpus. Let $x = [x_1, x_2, \ldots, x_Q]$ be the sequence of XLPC cepstrum describing source speaker. Similarly, let $y = [y_1, y_2, \ldots, y_R]$ be the sequence of XLPC cepstrum

describing target speaker. The weight, mean vector and covariance matrix of each component are estimated independently from the clusters obtained by VQ of the joint vectors c_n. The alignment is obtained by using dynamic time warping algorithm. The mean vector and covariance vectors are given by, $\mu_m^c = \begin{bmatrix} \mu_m^x \\ \mu_m^y \end{bmatrix}, \Sigma_m^c = \begin{bmatrix} \Sigma_m^{xx} & \Sigma_m^{xy} \\ \Sigma_m^{yx} & \Sigma_m^{yy} \end{bmatrix}$ where, μ_m^x and μ_m^y are the mean vectors of the m^{th} mixture component for source and target, respectively. Matrices Σ_m^{xx} and Σ_m^{yy} are the covariance matrices and Σ_m^{xy} and Σ_m^{yx} are cross-covariance matrices of the m^{th} mixture component for source and target speakers' respectively. The GMM is trained with an expectation-maximization (EM) algorithm using joint vectors.

The conversion function is implemented as a mixture of source and target XLPC cepstrum weighted by posterior probabilities of GMM [137]. The optimization of the conversion function is simplified by constraining both the covariance, Σ_m^{xx} & Σ_m^{yy} and cross-covariance matrices, Σ_m^{xy} & Σ_m^{yx} to be diagonal. Empirically, the number of GMM components, chosen, is three. The conversion is performed based on the minimum mean-square error as,

$$\begin{aligned} \widehat{y_n} &= E[y_n|x_n] \\ &= \sum_{m=1}^{M} P(m|x_n, \lambda^c) E_{m,n}^y \end{aligned} \tag{6.11}$$

where,

$$P(m|x_n, \lambda^c) = \frac{w_m \mathcal{N}(x_n; \mu_m^x, \Sigma_m^{xx})}{\sum_{i=1}^{M} w_i \mathcal{N}(x_n; \mu_i^x, \Sigma_i^{xx})} \tag{6.12}$$

and the mean vector $E_{m,n}^y$ can be written as,

$$E_{m,n}^y = \mu_m^y + \Sigma_m^{yx} \Sigma_m^{xx^{-1}} (x_n - \mu_m^x) \tag{6.13}$$

Here, $E_{m,n}^y$ represents the expectation value and $\widehat{y_n}$ is the converted target XLPC cepstrum vector. For each of the mixture, the conditional mean target vector given the mean source vector is calculated by linear conversion using the correlation between source and target features. As a result, the converted vector is expressed as the weighted sum of conditional mean vectors. The conditional probability that the source vector belongs to each one of the mixtures is used as weights. Further, the transformed XLP cepstrum $\widehat{y_n}$ is converted into XLP coefficients $\widehat{d_k}$. The modified XLP filter is obtained by using

the transformed XLP coefficient $\widehat{d}_k$, given by,

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} \widehat{d}_k z^{-k}} \tag{6.14}$$

here, $H(z)$ is the XLP filter predicted from p^{th} order XLP analysis, where p is chosen as $p = \frac{fs}{1000} + 4$. To generate the spectrally transformed vowel phonation, the unmodified XLP residual of the hypernasal speech is filtered using the modified all-pole filter $H(z)$. The spectrally enhanced speech signal is computed using the transfer function given as

$$V(z) = \frac{R(z)}{1 + \sum_{k=1}^{p} \widehat{d}_k z^{-k}} \tag{6.15}$$

In order to obtain both temporally and spectrally modified signal the modified XLP residual is passed through $H(z)$. The transfer function in Z domain can be represented as,

$$S_{tf}(z) = \frac{R'(z)}{1 + \sum_{k=1}^{p} \widehat{d}_k z^{-k}} \tag{6.16}$$

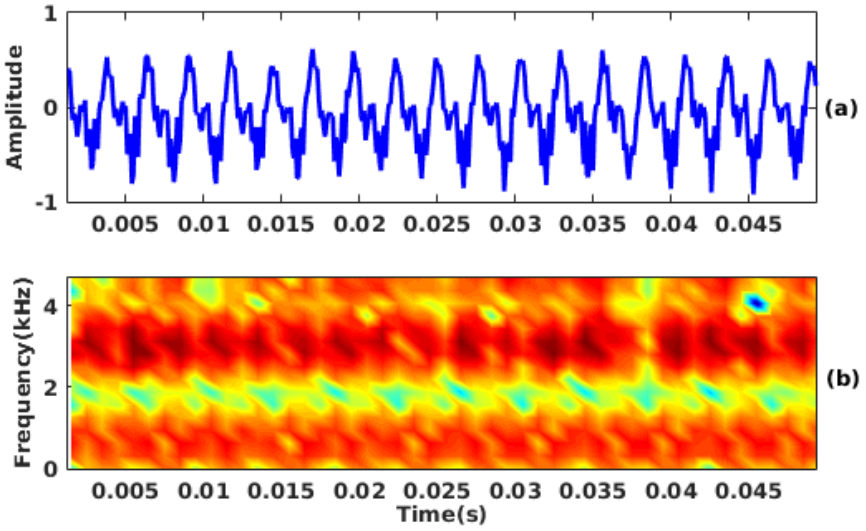

Figure 6.8: A segment of vowel /i/ phonation and corresponding spectrogram plot of modified hypernasal /i/ vowel phonation.

The speech signal and spectrogram plot of a modified hypernasal vowel are presented in Figure 6.8. A modified hypernasal spectrogram shown in Figure 6.8 (b) depicts more prominent resonances compared to the unprocessed hypernasal spectrogram in Figure 6.4 (d). The spectral energy level in Figure 6.8 (b) is quite closer to the non-CLP speaker vowel /i/ phonation shown in Figure 6.4 (b).

6.4 Experimental observations

The GMM based modification approach implemented for dysarthric speech enhancement in [1] is used for a comparative study. For simplification purposes, the system is referred as GMM-Dys throughout the chapter. Accordingly, the subjective and objective analysis are performed for the modified signal implemented using GMM based modification system proposed for dysarthric speech enhancement.

6.4.1 Objective evaluation

The modified hypernasal speech is assessed using three objective measures. The objective measures are evaluated as follows:

Pathological short-time objective intelligibility and pathological extended short-time objective intelligibility

The pathological short-time objective intelligibility (P-STOI) and pathological extended short-time objective intelligibility (P-ESTOI) are employed as our objective intelligibility measure to assess CLP speech intelligibility [267]. P-STOI is used to measure the impact of temporal distortions in CLP

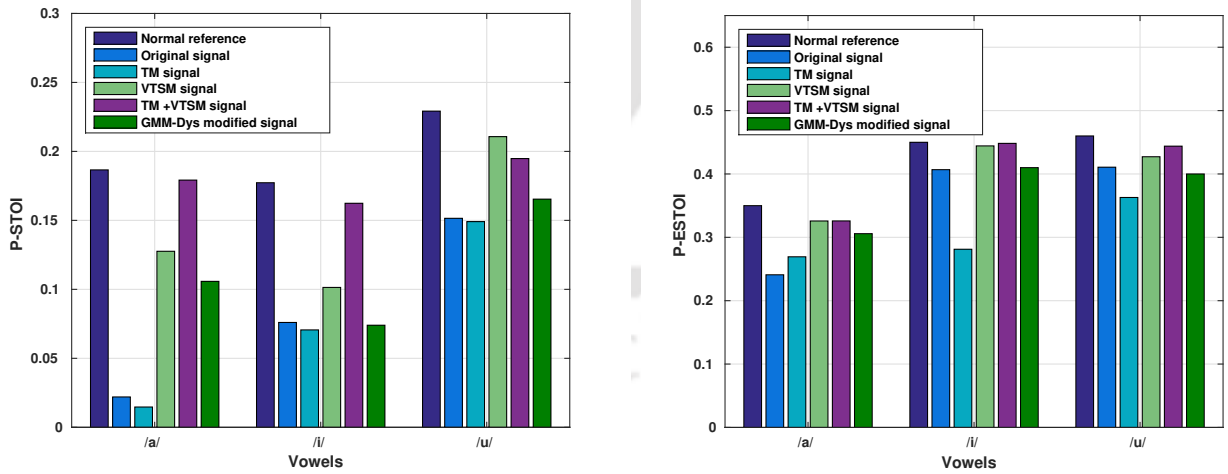

Figure 6.9: Bar plot showing P-STOI and P-ESTOI for three vowels, /a/, /i/ and /u/. TM & VTSM denotes XLP residual & vocal tract system characteristics modified hypernasal vowel, respectively. GMM-Dys refer to the signal processed using the GMM based voice conversion employed for dysarthric speech enhancement [1].

speech intelligibility, whereas, P-ESTOI analyze the impact of spectral distortions. The objective measures are based on comparing time-aligned hypernasal and non-CLP (target) signals. Before com-

puting, P-STOI, and P-ESTOI, first reference templates are created from the non-CLP speakers. In the current study, vowel-specific and gender-specific reference representations are used. One-third octave band analysis is applied to the time-frequency representation of non-CLP and hypernasal speech signals to estimate the objective measures.

The results of the P-STOI measure are depicted in Figure 6.9. The P-STOI measure shows that temporal and vocal tract system modified signal has intelligibility higher than the original hypernasal, and other variants of modified signals. TM denotes XLP residual modification and VTSM denotes vocal tract system modification. Comparatively, the GMM-Dys based modified signal has lower intelligibility than the combined temporal and vocal tract system modified (TM + VTSM) signal. From the P-STOI measure, it is also observed that the temporal and vocal tract system modified signal has STOI comparable to non-CLP reference STOI for vowel /a/ and /i/. However, the intelligibility of the combined temporal and vocal tract system modified signal of the vowel /u/ is slightly lower than the vocal tract system modified signal. In the case of P-ESTOI shown in Figure 6.9, a similar trend is followed, where the combined temporal and vocal tract system modified signal has comparable ESTOI value relative to other variants of the modified signal.

XLPC cepstral distortion

To measure the performance of spectral mapping, XLPC cepstral distortion (XCD) is used as another performance metric for objective evaluation. The XCD value here represents the mean of XCD values across all the frames. It is defined as a weighted Euclidean distance, which is expressed by,

$$\text{XCD (dB)} = \frac{1}{\text{N}} \sum_{n=1}^{N} 10 \times \log_{10} \sqrt{2 \sum_{i=1}^{I} (c_i - \hat{c}(i))^2} \tag{6.17}$$

where, n denotes the frame index, and N is the total number of frames. c_i and $\hat{c}_i$ denote the c^{th} target and converted cepstral coefficient respectively. The distortion measures are assessed for the four variants of each of the hypernasal vowels, denoted by $m1, m2, m3$, and $m4$. The distribution measure is also assessed for the GMM based modification system employed for dysarthric speech enhancement and it is denoted by $m5$. The four variants $m1$ through $m4$ corresponds to the original unmodified and three types of modified (TM, VTS, TM + VTS) vowels. The average XCD values for all the variants of the three vowels /a/, /i/ and /u/ are shown in Figure 6.10. It is observed that XCD values decreases from $m1$ to $m4$ for all the three vowels except for /u/ phonation where $m3$ is

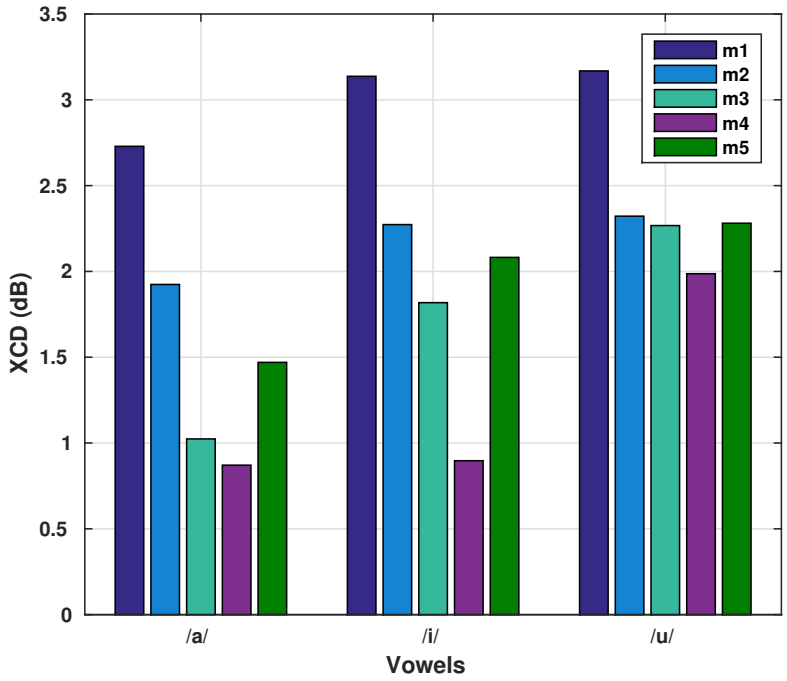

Figure 6.10: Bar plot showing XCD values for the three vowels where $m1$ denotes XCD value between non-CLP and unmodified hypernasal vowel, $m2$ denote XCD value between non-CLP and hypernasal vowel after XLP residual modification, $m3$ denote XCD value between non-CLP and hypernasal vowel after vocal tract system characteristics modification, $m4$ denote XCD value between target and hypernasal vowel after both the residual and vocal tract system characteristics modification, and $m5$ denote XCD value between target and hypernasal vowel modified using GMM-Dys.

comparable to $m2$. From Figure 6.10, it is also observed that XCD value is higher for $m5$ compared to $m3$ and $m4$. The decrease of the XCD values implies that the residual and vocal tract system modification reduces the spectral distortion of the vowels.

Support vector machine classifier

The support vector machine (SVM) is trained using the features extracted from the vowels of non-CLP speakers and that of CLP speakers. The classifier system is developed using radial basis function (RBF) kernel, and separate SVM models are built for each of the vowels /a/, /i/ and /u/, respectively. For testing, all the modified speech samples are used. The optimum values of the parameters, $\mathscr{C}$, and γ, are experimentally determined using the grid search method. During classification all the combinations of an RBF kernel are used in the range of $\mathscr{C} = [2^{-10}, 2^{-8}, \ldots, 2^{+8}, 2^{+10}]$ and $\gamma = [2^{-10}, 2^{-8}, \ldots, 2^{+8}, 2^{+10}]$. The best accuracy obtained in the considered range of $\mathscr{C}$ and γ is reported as the classification results. The performance of an SVM classifier for the non-CLP, hypernasal, and modified hypernasal vowels are tabulated in Table 6.1 and Table 6.2 using MFCC and XLPCC features, respectively.

Table 6.1: Accuracy (%) of vowel identification using the MFCC feature. Vowel subscripts o and m represent original and modified samples. GMM-Dys refer to the signal processed using the GMM based voice conversion employed for dysarthric speech enhancement.

Phoneme description	Hypernasal	Normal
$/a/_o$	71.13	28.87
$/a/_m$	13.88	86.12
$/a/_{GMM-Dys}$	48.33	51.67
$/i/_o$	92.00	8.00
$/i/_m$	49.20	50.80
$/i/_{GMM-Dys}$	47.68	52.32
$/u/_o$	79.80	20.20
$/u/_m$	32.42	67.58
$/u/_{GMM-Dys}$	45.33	54.67

Table 6.2: Accuracy (%) of vowels identification using the XLPCC feature. Vowel subscripts o and m represent original and modified samples. GMM-Dys refer to the signal processed using the GMM based voice conversion employed for dysarthric speech enhancement.

Phoneme description	Hypernasal	Normal
$/a/_o$	76.29	23.71
$/a/_m$	30.15	69.85
$/a/_{GMM-Dys}$	46.02	53.98
$/i/_o$	92.34	7.66
$/i/_m$	25.50	74.50
$/i/_{GMM-Dys}$	41.80	58.20
$/u/_o$	90.18	9.82
$/u/_m$	22.98	77.02
$/u/_{GMM-Dys}$	47.72	52.28

Significant improvements are observed for the transformed vowels compared to the original hypernasal vowels. An increase in the recognition rates of the transformed vowels tending towards non-CLP vowels is observed for both MFCC and XLPCC features.

The improvement in P-STOI and P-ESTOI, reduction in XCD values, and the recognition performance of an SVM classifier are observed to be different for all the three vowels. This may be due to the difference in the vowel characteristics like high, low, and mid vowel. As a result, specific vowels are more prone to nasalization. The nasalization in a particular vowel is effectively reduced using the enhancement approach described in Subsection 6.3.2. Thus for particular vowels, distortion due to

nasalization varies, and similarly, its reduction also varies based on the level of nasality.

6.4.2 Subjective evaluation

Listening experiments are conducted to evaluate the speech intelligibility of hypernasal speech. The modified hypernasal speech is synthesized using enhanced XLP residual and vocal tract system characteristics outlined in Section 6.3. The listening test aims to determine the hypernasality effect in the modified hypernasal vowels relative to that of unmodified hypernasal vowels. Three vowels (/a/, /i/, /u/) in /CV/ structures are considered as a test material for listening experiments. A total of 10 normal-hearing naive listeners have participated in the listening test. The listeners are research scholars bearing knowledge of speech technology. Speech utterances were played to the listeners through headphones at a comfortable listening level. Each listener can control the hearing level according to their comfortability. A total of 75 speech samples (original + speech after XLP residual signal modification+ vocal tract system modification + combination of residual and vocal tract system modification + GMM-Dys modified signal, representing 5 conditions for 5 sets of each of the 3 vowels, $5 \times 5 \times 3 = 75$) are provided to the listeners for transcription and deriving MOS.

At first, the intelligibility of a modified hypernasal speech signal is measured by having the participants transcribe what they hear from the speech signal. As this work deals with vowel modification, the listeners were instructed to concentrate on vowels only. Therefore, the listeners attempt to identify the vowels in the /CV/ structures. The original hypernasal vowel and four variants (TM, VTS, TM+VTS, GMM-Dys) of the modified signal are presented to the listeners. The participants are told to ignore those /CV/ structures, which they cannot discern.

Table 6.3: Performance in phoneme accuracy (%) for hypernasal vowels by naive listeners. TM & VTSM denotes XLP residual & vocal tract system characteristics modified hypernasal vowel, respectively. GMM-Dys refer to the signal processed using the GMM based voice conversion employed for dysarthric speech enhancement.

Phonemes	Original	TM	VTSM	TM + VTSM	GMM-Dys
/a/	44.56	46.89	78.32	81.11	57.80
/i/	40.21	48.60	84.51	88.23	59.36
/u/	43.50	45.00	75.54	80.64	54.35

Table 6.3 shows the percentage of vowels correctly identified by the listeners relative to the expected vowel. The values reported in Table 6.3 are the average values computed across the listeners. From

Table 6.3, it is observed that the TM + VTSM signal has higher intelligibility than other variants of the modified signal. Considering all the original vowels, the vowel /i/ is observed to have the least recognition accuracy compared to vowel /a/ and /u/. However, the vowel /i/ after TM + VTSM modification, shows maximum recognition accuracy relative to vowel /a/ and /u/. The possible reason for this observation may be attributed to the vowel /i/ being a high vowel, which is more prone to nasalization.

The subjective evaluation is also carried out using mean opinion score (MOS) for all the three vowels, respectively. The listeners were asked to give a score against each of speech files ranging between 1 to 4 based on nasalization, where 4 corresponds to most nasalized speech and 1 for the least. Before the listening test, all the listeners are provided with severe hypernasal and non-CLP

Table 6.4: Subjective evaluation for mean opinion score by naive listeners. TM and VTSM denote XLP residual and vocal tract system characteristics modified hypernasal speech, respectively. GMM-Dys refer to the signal processed using the GMM based voice conversion employed for dysarthric speech enhancement.

Phonemes	Original	TM	VTSM	TM + VTSM	GMM-Dys
/a/	2.77	2.72	2.10	1.97	2.21
/i/	3.46	3.30	2.00	1.50	2.50
/u/	3.10	2.80	2.32	1.39	2.26

speech signals to familiarize them with the expected target speech signal and the disordered hypernasal speech signal. The order of the speech samples played to the listeners is randomized to avoid any bias towards any signal. Table 6.4 shows the averaged scores corresponding to each method. In the case of vowel /a/, TM + VTSM gives better performance compared to original and other modified signals. However, the nasality rating of unprocessed hypernasal vowel /a/ is observed to be lower compared to vowels /i/ and /u/. It may be because, /a/ is a low vowel, and it is least affected by hypernasality among the three vowels considered in this study. As a result, reduction in nasality scores over TM, VTSM, TM + VTSM, and GMM-Dys signal is relatively lower. Considering vowel /i/, a significant reduction in nasality rating score is observed for all the three variants of enhancement and GMM-Dys modified signal. The vowel /i/ being a high vowel is observed to possess maximum nasality score among the three studied vowels. In the case of mid vowel /u/, the nasality rating score and relative reductions lie in between vowel /i/ and /a/. Overall, MOS values reflect that TM + VTSM signals

give better performance for all the three vowels than TM, VTSM, and GMM-Dys, respectively.

6.5 Summary

The present work described three variants of enhancement, XLP residual modification, vocal tract system characteristics modification, and combined modification of XLP residual and vocal tract system characteristics to enhance the hypernasal speech [234]. An XLP method is used to parameterize the residual and vocal tract system components of the hypernasal speech signal. Modification of the XLP residual is carried out by using a fine weight function convolved around the significant GCI locations extracted using a ZFF. The vocal tract system characteristics are modified spectrally using GMM based spectral conversion trained on the non-CLP and hypernasal speech data. The performance evaluation is conducted using different objective measures, namely, P-STOI, P-ESTOI, XCD, SVM classifier, and two types of subjective measures, transcription test, and MOS. The results from subjective and objective evaluations indicate that out of the three variants of hypernasal speech enhancement, combined modification of XLP residual signal and vocal tract system components gives better speech enhancement compared to either XLP residual or vocal tract system characteristics modification alone. Consequently, the enhanced speech signal output has reduced the nasalization effect compared to the unmodified original speech.

7

Combined framework for the word-level intelligibility enhancement

Publications

[268] Protima Nomo Sudro, S. R. Mahadeva Prasanna,"Processing Phoneme Specific Segments for Cleft Lip and Palate Speech Enhancement", *APSIPA* 2021.

Contents

Overview *The speech intelligibility of the cleft lip and palate (CLP) speakers is distorted due to the deformation in their articulatory system. For addressing the same, the works in the earlier chapters perform phoneme-specific modification in consonant-vowel-consonant-vowel (CVCV) structures. In CLP speech, both the articulation error and the nasalization distorts the intelligibility of a word. Modification of a specific phoneme may not always yield an enhanced word-level intelligibility. Accordingly, all the distorted phonemes in a word must be modified to achieved word-level intelligibility improvement. For such cases, it is important to identify and isolate the phoneme specific error based on the knowledge of specific speech distortions. Motivated by that, in this work, some of salient phoneme specific enhancement approaches discussed in the earlier chapters are combined. Further, we demonstrate their effectiveness in improving the word-level intelligibility of CLP speech. The enhanced speech samples are evaluated using subjective and objective evaluation metrics.*

7.1 Introduction

In the earlier chapters, the focus is made on the (a) segmentation of misarticulated fricative followed by modification of the same, (b) development of approaches for the detection of burst event in stop consonants followed by transforming the misarticulated stops, and (c) modification of vowels using spectral and temporal processing. The first enhancement work addresses three types of fricative /s/ misarticulation: palatalized /s/, phoneme specific nasal air emission (PSNAE) distorted /s/ and glottal stop substituted /s/. The misarticulated fricatives are detected automatically in a /CVCV/ structure, then the detected segments are modified using spectral compression, spectral tilt modification, and insertion method. The second work deals with the detection and modification of three misarticulated stops: /k/, /t/, and /T/. The velar stop /k/ is substituted by a glottal stop sound. Three types of misarticulations of alveolar /t/ and retroflex /T/ are studied: glottal stop, velar, and palatalized substitutions. The stop modification is performed using NMF based method in a /CVCV/ structure. The third work on CLP speech enhancement task consists of nasalized vowel /a/, /i/, and /u/ modifications in a /CV/ structure. The vowel modifications are carried out using GMM based spectral conversion method and temporal processing using a weighting function.

In the CLP speech analysis and enhancement tasks performed in clinical settings, the SLPs first work on isolated phonemes. Once a CLP speaker has mastered the correct phoneme production, the SLPs embed the phoneme in a word and analyze the speech intelligibility of an entire word. Further,

this process is observed in a sentence, short phrase and conversational speech. In [8], it is reported that a speaker may produce a phoneme correctly in isolation, but the same may be produced in error in a word-level and sentence-level due to the influence of phonetic contexts. Hence, the SLPs evaluate and attempt to enhance the CLP speech by minimizing the influence of phonetic contexts. In the similar direction, the present work also focuses on the word-level intelligibility via combination of phoneme specific modification techniques. The previous chapter works showed an ability to modify the phoneme specific distortions for fricative /s/, stop consonant /k/, /t/, /T/ and vowel /a/, /i/, and /u/ phonations. In the present chapter, the relevant phoneme specific enhancement techniques are combined to improve the entire word-level intelligibility. Here, the considered /CVCV/ words happen to be the possible combinations of the phonemes studied in the previous chapters.

7.2 Transforming word-level speech intelligibility

In a /CVCV/ word, when a specific transformation method is used to modify the entire word, it is observed that the speech sounds muffled. Thus, further processing is required to achieve a good quality enhanced speech. Several issues exist in performing the entire word-level intelligibility because different phonemes in a word may exhibit different type of speech distortions along with co-articulation impact. In a word, some of the phonemes may get nasalized, some are substituted by nasal consonants, and various types of other misarticulations affect the CLP speech intelligibility and quality. It is challenging to detect such misarticulations in an unsupervised method. Therefore, certain assumptions are made prior to the enhancement task.

The important acoustic-phonetic cues related to obstruents are degraded due to the production error. The cues, namely, transient burst, frication noise, and formant dynamics in the adjacent transition region mostly gets degraded. Therefore, it is assumed that the deviations in their productions that are reflected on the acoustic signal may be correlated with the perceived CLP speech intelligibility. On the other hand, due to nasalization, the voiced sounds are mostly affected. In a /CVCV/ word, when the obstruent is modified, distorted voiced sounds having co-articulation could still affect the overall intelligibility. This renders the modification of the voiced sounds with nasalization and co-articulation. Hence, characterizing all the deviations using relevant acoustic features followed by modification of the same are expected to yield intelligible speech. The knowledge of deviated acoustic characteristics explored for the analysis of CLP speech in the previous chapters can be used to enhance

the distorted word-level intelligibility. Moreover, the methods corresponding to the segmentation of misarticualted fricative /s/ and the stop consonants reported in Chapter 4 and Chapter 5 are exploited in this chapter for the segmentation of obstruents and vowels in the word.

7.2.1 Transformation of fricative-vowel-fricative-vowel words

In this subsection, the study is carried out using three isolated words (/sasa/, /sisi/, and /susu/) in the database which happen to be the combination of fricative /s/ and vowels /a/, /i/, and /u/. The details of these three words are described in Table 3.11 of Chapter 3. The analysis, segmentation, and modification of fricatives in fricative-vowel-fricative-vowel (FVFV) words are performed using the techniques exploited in Chapter 4. The vowels in the /FVFV/ words are enhanced using spectral and temporal processing methods described in Chapter 6.

The approaches exploited for fricative and vowel modification corresponds to spectral compression, insertion, spectral conversion and temporal processing. If spectral compression technique is to be applied to modify the entire word /sasa/ or /sisi/ or /susu/, then lower frequency region compression will yield enhanced fricatives but it will deteriorate the vowels at the same time because low frequency energy is important for vowel perception [232]. In certain cases, if vowels are nasalized, insertion method may be applicable but its perceptual quality may get distorted as it might not preserve the individuality information of the CLP speaker. Considering the temporal processing method described in Chapter 6, a weighting function is applied around the GCI locations to emphasize the significant GCI events and suppressed any interfering signal components around it. If the /FVFV/ words are processed using temporal processing method, it is speculated that it might not result in effective transformation. The reason may be attributed to the fact that the excitation source for fricative is a noise source signal and with temporal processing important signal components might get suppressed. As each of the phonemes have different spectral and temporal phenomena, the distortion caused by the articulatory impairment affects each of the phoneme differently. Further, considering the spectral conversion method, mapping the distorted CLP spectra to a desired target speech spectra is a good strategy towards attaining CLP speech enhancement. Hence, an attempt is made to enhance /FVFV/ words using GMM based spectral conversion method.

The enhanced /sasa/ word using GMM based spectral conversion is depicted in Figure 7.1 (e)-(f). After performing enhancement, it is observed that GMM based spectral conversion of hypernasal speech results in muffled speech. The analysis of the modified speech signal showed that the spectrally

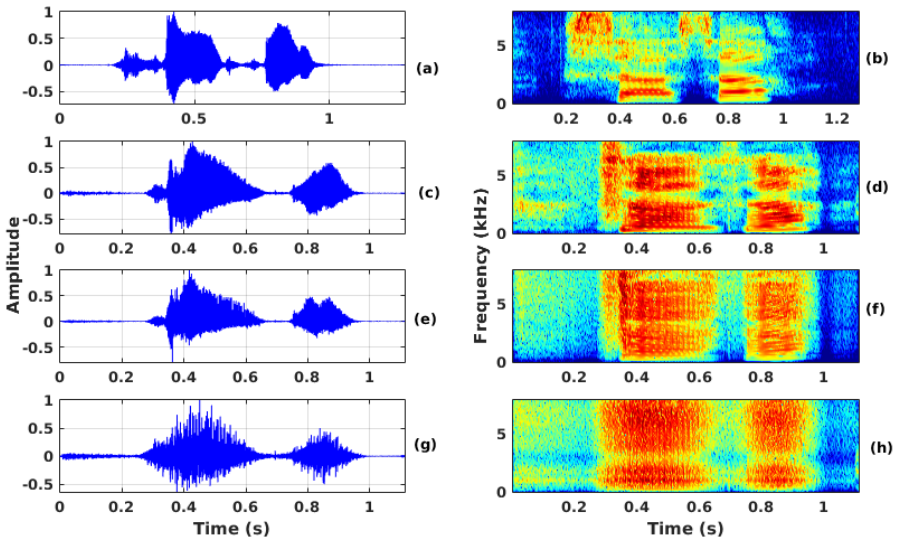

Figure 7.1: Comparison of (a)-(b) non-CLP (healthy) /sasa/, (c)-(d) NAE distorted /sasa/, (e)-(f) enhanced /sasa/ using GMM based spectral conversion method and (g)-(h) enhanced /sasa/ using NMF based spectral conversion method .

modified speech is observed to have low nasalization compared to original /FVFV/ word depicted in Figure 7.1 (c)-(d), but the fricatives are ambiguous and the overall speech quality is still degraded. The speech degradation may be attributed to the oversmoothing effect. Because of the oversmoothing effect the formants of the vowels are observed to have larger bandwidth, smaller peak-to-valley ratio, and the fricative spectrum are observed to have deviant acoustic characteristics and spectral tilt. Several techniques are proposed in the literature to overcome the smoothening affect, where one of the approach corresponds to the parametric voice conversion method which is reported to alleviate the oversmoothing effect. Therefore, NMF based speech enhancement is used to modify the /FVFV/ word and shown in Figure 7.1 (g)-(h). The modified /FVFV/ word shows some improvement compared to original /FVFV/ word but its acoustic characteristics are still degraded compared to that of the non-CLP (healthy) /FVFV/ word depicted in Figure 7.1 (a)-(b). Although some improvements are observed, such as reduction in nasalization and spectral prominence in the high-frequency region for fricatives. However, further observation showed that low-frequency energy in the fricatives remain and the formants of the vowel are not distinct. Thus, it projects the importance of processing different class of sound units separately utilizing the phoneme-specific knowledge.

7.2.2 Transformation of consonant-vowel-consonant-vowel words

In this part of the study, the words, namely, /kaka/, /kiki/, /kuku/, /tata/, /titi/, /tutu/, /TaTa/, /TiTi/, and /TuTu/ are addressed. These words consists of the combination of stops /k/, /t/ and /T/ in the vowel contexts /a/, /i/, and /u/, respectively. The details of these /CVCV/ words are described in Table 3.11 of Chapter 3. Prior to the modification of the /CVCV/ words, the analysis, detection, and modification of burst of the stop consonants are carried out using the method discussed in Subsection 5.2.1 of Chapter 5. The vowel analysis and modification of the same is performed using the approaches reported in Subsection 6.3 of Chapter 6.

Considering the reported phoneme-specific enhancement methods in the earlier chapters, their feasibility in the modification of /CVCV/ words are first analyzed. The spectral compression technique may not effectively transform the /CVCV/ word because the spectral prominence of consonants are not confined to a specific frequency band. For example, the velar stop /k/ is characterized by prominent spectral energy in the low-frequency region, whereas, the alveolar stop /t/ shows spectral prominence around mid-frequency region and retroflex /T/ shows spectral prominence above 2 kHz. Additionally, formants of the vowel are observed in the low-frequency region. Therefore, spectral energy compression in the low-frequency region will result in further deterioration of the /CVCV/ words. Temporal processing of the /CVCV/ word will result in vowel modification only. By using insertion method, artificially synthesized phoneme can be used for speech modification. However, insertion method does not preserve the individuality information. Hence, spectral compression, temporal processing and insertion method might not effectively enhance the entire /CVCV/ word.

The spectral prominence for consonants ranges from low to high-frequency region and in vowels the lower frequency region are mostly considered to carry important perceptual information. Hence, the modification technique designed for a specific category of phoneme may or may not be effective in transforming the disordered nature of all the phonemes in a word. As stated in the previous section, mapping the disordered speech spectra into that of the non-CLP (target) speech spectra may improve the speech intelligibility and quality. Hence, the modification of /CVCV/ word is also attempted using GMM and NMF based spectral conversion method, respectively.

For illustration, the transformed signals for the word /kaka/ is shown in Figure 7.2. From the figure, it is observed that the vowel formants and the consonants in Figure 7.2 (e)-(f) and Figure 7.2 (g)-(h) are not close to non-CLP speech characteristics shown in Figure 7.2 (a)-(b). Based on the above

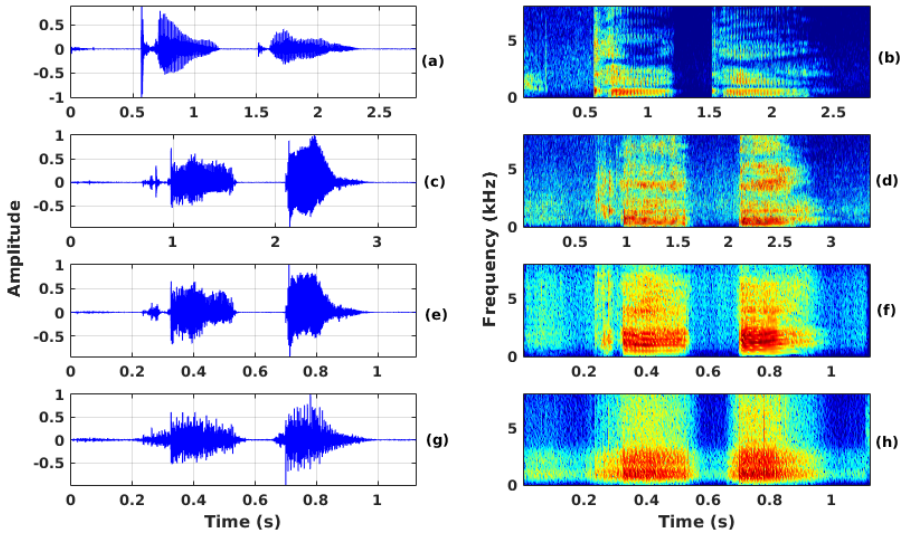

Figure 7.2: Comparison of (a)-(b) non-CLP /kaka/, (c)-(d) misarticulated /kaka/, (e)-(f) enhanced /kaka/ using GMM based spectral conversion method and (g)-(h) enhanced /kaka/ using NMF based spectral conversion method.

figures, it is observed the enhanced speech signal does not show significant improvement relative to the original unprocessed signal. The reason may be attributed to the complex relationship of the misarticulations and nasality in CLP speech. Hypernasality and articulation error both show an impact in the same word reducing the speech intelligibility and quality both. In the above figures, for an illustration only /kaka/ is depicted. However, similar analysis are observed for other /CVCV/ word misarticulations, for example, /kiki/, /kuku/, /tata/, /titi/, /tutu/, /TaTa/, /TiTi/, and /TuTu/. Both the NMF and GMM based spectral conversion had shown some improvement in the speech characteristics, however, they are not able to effectively enhanced the speech signal as close to non-CLP. Therefore, this necessitates further analysis of the signal characteristics and then perform modification. Hence, phoneme specific enhancement can be attempted to observe the impact on the overall speech intelligibility and quality of a word.

7.3 Experimental evaluation

In this section, the impact of independent modification of all the phonemes in a word are analyzed. From Figure 7.1 and Figure 7.2, it is observed that when the entire word is modified using a specific modification technique, the different attributes of the speech distortions in all the phonemes are not

addressed effectively. Due to the impairment, both articulation error and nasalization are observed in

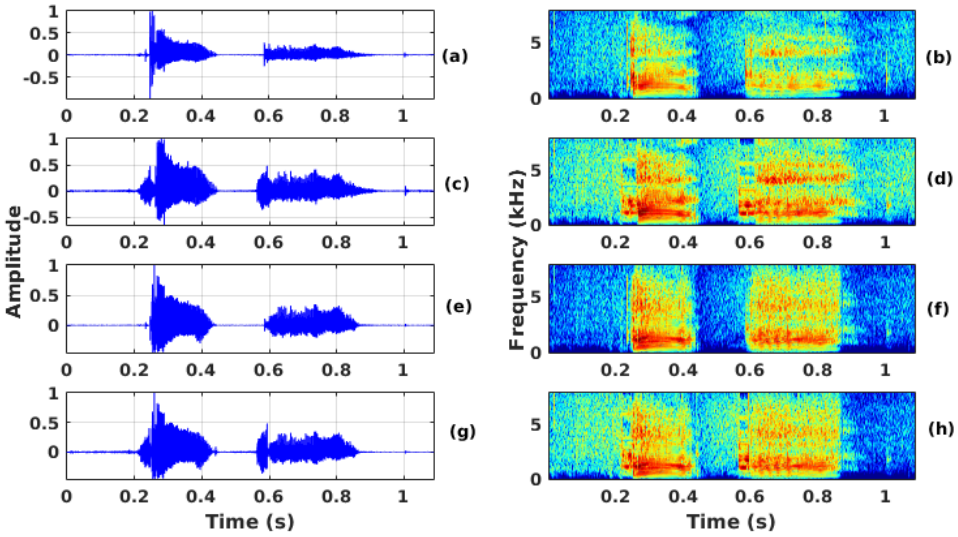

Figure 7.3: Illustration of waveform and spectrogram of: (a)-(b) clp /kaka/, (c)-(d) modified /k/, (e)-(f) modified /a/, and (g)-(h) modified /k/ and /a/.

CLP speech. When each of the phonemes in a word are modified independently as shown in Figure 7.3 and Figure 7.4, a significant modification in the articulation error and nasalization is observed. To

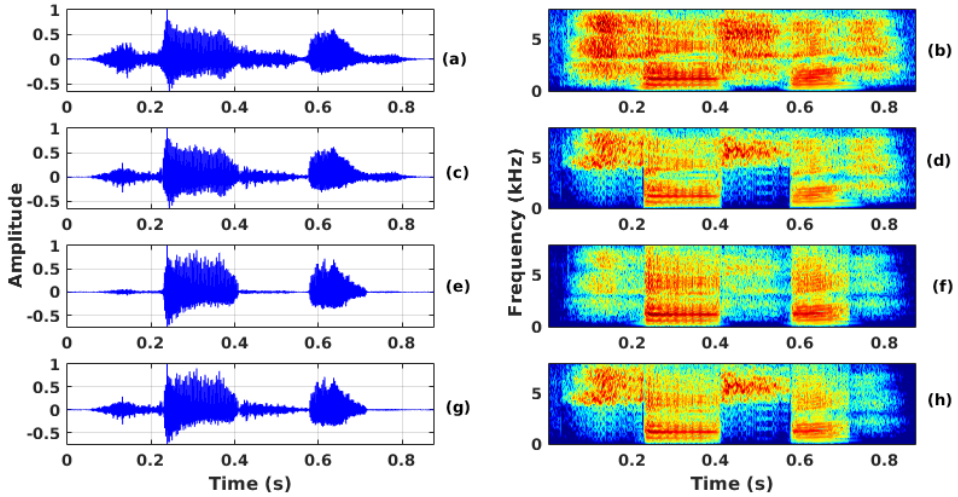

Figure 7.4: Illustration of waveform and spectrogram of: (a)-(b) clp /sasa/, (c)-(d) modified /s/, (e)-(f) modified /a/, and (g)-(h) modified /s/ and /a/.

have an enhancement system capable of performing the desired task, it is essential to improve all the speech distortions of a word. The impact of the word-level speech enhancement is analyzed in three

Table 7.1: P-STOI values for different combination of nonsensical words. F denote fricative /s/, C_1 denote consonant /k/, C_2 denote consonant /t/, C_3 denote consonant /T/, GS denote glottal stop substitution, PA denote palatalized articulation, PSNAE denote phoneme specific nasal air emission, and velar denote velar substitutions.

Speech samples	Normal Reference	$CLP_{original}$	$CLP_{modified}$		
			Obstruent	Vowel	Obstruent + vowel
/FVFV/ (GS)	0.88	0.02	0.27	0.23	0.46
/FVFV/ (PA)	0.88	0.11	0.26	0.22	0.44
/FVFV/ (PSNAE)	0.88	0.12	0.35	0.21	0.49
/C_1VC_1V/ (GS)	0.84	0.04	0.11	0.33	0.40
/C_2VC_2V/ (GS)	0.81	0.12	0.14	0.16	0.37
/C_2VC_2V/ (Velar)	0.81	0.13	0.14	0.18	0.39
/C_2VC_2V/ (PA)	0.81	0.13	0.14	0.19	0.42
/C_3VC_3V/ (GS)	0.82	0.15	0.17	0.24	0.43
/C_3VC_3V/ (Velar)	0.82	0.18	0.19	0.23	0.39
/C_3VC_3V/ (PA)	0.82	0.16	0.18	0.18	0.42

Table 7.2: P-ESTOI values for different combination of nonsensical words. F denote fricative /s/, C_1 denote consonant /k/, C_2 denote consonant /t/, C_3 denote consonant /T/, GS denote glottal stop substitution, PA denote palatalized articulation, PSNAE denote phoneme specific nasal air emission, and velar denote velar substitutions.

Speech samples	Normal Reference	$CLP_{original}$	$CLP_{modified}$		
			Obstruent	Vowel	Obstruent + vowel
/FVFV/ (GS)	0.25	0.02	0.06	0.13	0.18
/FVFV/ (PA)	0.25	0.06	0.09	0.16	0.22
/FVFV/ (PSNAE)	0.25	0.01	0.02	0.15	0.24
/C_1VC_1V/ (GS)	0.19	0.03	0.07	0.16	0.20
/C_2VC_2V/ (GS)	0.31	0.09	0.15	0.17	0.23
/C_2VC_2V/ (Velar)	0.31	0.07	0.08	0.10	0.16
/C_2VC_2V/ (PA)	0.31	0.06	0.10	0.17	0.23
/C_3VC_3V/ (GS)	0.29	0.12	0.17	0.20	0.22
/C_3VC_3V/ (Velar)	0.29	0.15	0.15	0.21	0.19
/C_3VC_3V/ (PA)	0.29	0.01	0.07	0.11	0.27

stages. At first the impact on word-level intelligibility is evaluated for the enhanced speech obtained by transforming the obstruents only. In the second step, the enhanced speech obtained using only vowel modification is evaluated. Finally, in the third step, the enhanced speech obtained by modifying both the obstruent and vowels are evaluated.

7.3.1 Objective evaluation

The modified CLP speech is assessed using pathological short-time objective intelligibility (P-STOI) and pathological extended short-time objective intelligibility (P-ESTOI) measures [267]. Automatic speech recognition (ASR) and mel cepstral distortion (MCD) are also used as the objective metrics.

The objective measures corresponding to P-STOI, P-ESTOI, ASR and MCD are noted in Ta-

ble. 7.1, Table. 7.2, Table. 7.3, and Table. 7.4, respectively. The P-STOI and P-ESTOI values are computed for time-aligned CLP speech signals and non-CLP (reference) signals. Before comparing the objective measures, word specific reference templates are created from non-CLP speech. The reported values are obtained by averaging all measures across all the listeners corresponding to the specific errors in all the vowel contexts. P-STOI values shown in Table 7.1 indicate that compared to original misarticulated CLP speech, modification of any specific error (obstruents misarticulation or vowel nasalization) improves the intelligibility. However, from the P-STOI values, it is also observed that modification of obstruents misarticulation and vowel nasalization both provide higher P-STOI values as compared to standalone modification of either obstruents or vowels. Similar observations are noted for the P-ESTOI values tabulated in Table 7.2. In certain cases, the objective intelligibility values of combined modifications are observed to be comparable with standalone modification. The probable reason may be attributed to the fact that either obstruent or vowel in the word is less distorted, hence resulting in comparable values.

For an illustration, the original and modified CLP words are analyzed using ASR performance. As the speech data for this study are in the Kannada language, a Kannada ASR system is developed using KALDI speech recognition toolkit [205]. The ASR system performance for various phoneme modification categories is measured using phone error rate (PER) metric. The PER for the original and modified CLP speech is shown in Table 7.3.

Table 7.3: Phoneme error rate (%) for various combination of nonsensical words. F corresponds to fricative /s/, C_1 corresponds to consonant /k/, C_2 corresponds to consonant /t/, C_3 corresponds to consonant /T/, GS corresponds to glottal stop substitution, PA corresponds to palatalized articulation, PSNAE corresponds to phoneme specific nasal air emission, and velar corresponds to velar substitutions.

Speech samples	$CLP_{original}$	$CLP_{modified}$		
		Obstruent	Vowel	Obstruent + vowel
/FVFV/ (GS)	68.85	56.08	58.15	54.53
/FVFV/ (PA)	59.23	38.80	36.83	31.58
/FVFV/ (PSNAE)	63.38	39.99	35.28	32.97
/C_1VC_1V/ (GS)	67.17	55.44	58.78	53.14
/C_2VC_2V/ (GS)	64.30	53.04	59.27	52.73
/C_2VC_2V/ (Velar)	68.46	59.96	50.18	54.04
/C_2VC_2V/ (PA)	49.72	43.52	48.06	42.98
/C_3VC_3V/ (GS)	77.31	63.09	65.92	62.28
/C_3VC_3V/ (Velar)	75.26	64.91	67.22	63.29
/C_3VC_3V/ (PA)	72.54	61.43	62.18	57.65

Compared to the original unprocessed CLP speech, the recognition performance of the modified speech is relatively higher. The modification of a specific phoneme also yield reduced PER value

compared to that of the original distorted CLP speech. However, better performances are observed when both the phonemes in the word are modified. In certain instances, the PER of the combined modification is comparable to the standalone modification of the phoneme. This implies that, in those utterances, the distortion caused by that specific phoneme is dominant.

Table 7.4: MCD values for different combination of nonsensical words. F corresponds to fricative /s/, C_1 corresponds to consonant /k/, C_2 corresponds to consonant /t/, C_3 corresponds to consonant /T/, GS corresponds to glottal stop substitution, PA corresponds to palatalized articulation, PSNAE corresponds to phoneme specific nasal air emission, and velar corresponds to velar substitutions.

Speech samples	$CLP_{original}$	$CLP_{modified}$		
		Obstruent	Vowel	Obstruent + vowel
/FVFV/ (GS)	13.00	11.20	11.80	11.90
/FVFV/ (PA)	12.94	11.31	12.23	12.29
/FVFV/ (PSNAE)	13.00	11.50	12.18	12.26
/C_1VC_1V/ (GS)	12.34	11.91	11.41	11.53
/C_2VC_2V/ (GS)	12.73	10.81	11.91	11.98
/C_2VC_2V/ (Velar)	13.58	10.91	11.75	12.22
/C_2VC_2V/ (PA)	13.58	10.58	11.90	12.13
/C_3VC_3V/ (GS)	13.74	10.85	11.76	12.02
/C_3VC_3V/ (Velar)	13.80	10.67	11.87	12.27
/C_3VC_3V/ (PA)	13.91	10.46	11.88	12.36

MCD is computed using equation 4.8 expressed in Subsection 4.5.1 of Chapter 4. Considering the MCD values shown in Table 7.4, it is observed that the combined modification of the obstruent and vowels, results in lower MCD values relative to that of the original distorted CLP words. The MCD values reported in Table 7.4 are averaged across all the listeners corresponding to the specific errors in all the vowel contexts. The lower MCD values of the modified words indicate that the spectral difference between non-CLP and misarticulated words are reduced significantly for all the words evaluated in the study.

The objective evaluation results signifies that different types of speech distortions influences each of the phonemes differently, resulting in less intelligible CLP speech. Hence, modification of each of the phonemes shows an improved performance compared to the isolated phoneme modifications alone.

7.3.2 Subjective evaluation

Listening experiment is also carried out to assess the word-level intelligibility. The modified CLP speech is evaluated to check its quality using mean opinion score (MOS). The MOS with a 5-point rating scale (1 = bad, 2 = fair, 3 = good, 4 = very good, and 5 = excellent) is used for speech quality evaluation. A total of 10 naive listeners have participated in the study and the speech samples were randomly numbered to avoid any bias towards the original or modified speech.

Table 7.5: Distribution of the speech samples presented to the listeners.

Speech samples	No. of words
/FVFV/	3 vowel contexts × 3 errors × 4 variations = 36
/C_1VC_1V/	3 vowel contexts × 1 error × 4 variations = 12
/C_2VC_2V/	3 vowel contexts × 3 errors × 4 variations = 36
/C_3VC_3V/	3 vowel contexts × 3 errors × 4 variations = 36
Total	= 120

Table 7.6: MOS for different combination of nonsensical words. F corresponds to fricative /s/, C_1 corresponds to consonant /k/, C_2 corresponds to consonant /t/, C_3 corresponds to consonant /T/, GS corresponds to glottal stop substitution, PA corresponds to palatalized articulation, PSNAE corresponds to phoneme specific nasal air emission, and velar corresponds to velar substitutions.

Speech samples	$CLP_{original}$	$CLP_{modified}$		
		Obstruent	Vowel	Obstruent + vowel
/FVFV/ (GS)	1.00±0.09	1.5±0.07	2.10±0.15	3.10±0.50
/FVFV/ (PA)	1.10±0.15	1.54±0.55	2.42±0.26	3.92±0.50
/FVFV/ (PSNAE)	1.50±0.40	1.90±0.60	2.50±0.20	3.80±0.20
/C_1VC_1V/ (GS)	1.58±1.17	1.88±0.75	2.85±0.46	3.72±0.45
/C_2VC_2V/ (GS)	1.12±1.11	1.53±0.51	2.27±0.47	3.57±0.99
/C_2VC_2V/ (Velar)	1.08±1.19	1.97±0.99	2.92±0.88	3.82±0.55
/C_2VC_2V/ (PA)	1.94±0.98	2.02±0.58	2.80±0.76	3.71±0.87
/C_3VC_3V/ (GS)	1.20±1.01	1.54±0.74	2.10±0.92	2.99±0.22
/C_3VC_3V/ (Velar)	1.15±0.83	1.72±1.20	2.30±0.57	3.05±1.09
/C_3VC_3V/ (PA)	1.53±1.25	1.72±0.58	2.70±0.89	3.59±1.12

The listeners bear the knowledge of speech science and technology. Each of the listener have evaluated 120 speech samples. The details of the number of words presented to the individuals are shown in Table 7.5. The MOS values are derived by averaging all the MOS values across each vowel context of the specific word from all the listeners. The averaged MOS values shown in Table 7.6 indicate that the combined modification of the obstruent and vowel yield significant improvement compared to the original and standalone modified CLP speech.

7.4 Summary

The word-level intelligibility is attempted by using the salient properties of phoneme-specific modifications discussed in the previous chapters. A comparison study is also done to observe whether the specific transformation method exploited in the above chapters can improve the entire word intelligibility. When a specific transformation method is used to modify the word as a whole, it is observed that the speech sounds are still distorted. Hence, phoneme-specific modifications are exploited to observe its impact in word intelligibility. From the evaluation results, it is observed that, improvement in the word-level intelligibility can be achieved when all the phonemes in a word are modified independently.

8

Conclusions

Contents

Overview

This chapter provides the summary and conclusions of the works presented in the thesis towards enhancing the articulation error and hypernasality in cleft lip and palate(CLP) speech. Based on the contributions and different investigations, few insights are discussed for future research.

8.1 Summary of the work

In this thesis, an attempt is made to analyze and modify misarticulation of obstruents (fricatives and stop consonants) and vowel nasalization. At first, the misarticulated fricative /s/ is automatically segmented and then subject to modification. In the next work, the compensatory errors produced for stop consonants are modified using nonnegative matrix factorization (NMF) method. Three unvoiced stop (/k/, /t/, /T/) misarticulations are addressed in the thesis. In several cases, individuals with CLP exhibit both the articulation error and the nasalization of voiced sounds. For such aspects, only the articulation error modification may not yield the desired speech intelligibility and quality. Hence, this renders the need for de-nasalizing the voiced sounds. Therefore, three vowels (/a/, /i/, and /u/) are analyzed and modified using temporal processing and Gaussian mixture model (GMM) based spectral conversion. Further, the combined impact of each of the phoneme specific modifications are evaluated for word-level intelligibility.

The incorporated contributions of the thesis are summarized as follows:

(i) **CLP Speech Database Development:** The unavailability of CLP speech database in public domain poses a major limitation in this regard. Therefore, it is necessary to first develop a database consisting of CLP speech.

The speech samples for this study are collected from AIISH, Mysuru, India. Both the CLP and non-CLP (healthy) speakers data are recorded in a sound proof room using a speech level meter (Bruel and Kjær) at a sampling frequency of 48 kHz and 16-bit resolution [173]. The speech stimuli consists of nonsensical words, meaningful words and short phrases. Forty-two CLP speakers (24 males and 18 females) in the age range of $7-12$ years participated in the study. Forty-two non-CLP speakers (20 males and 22 females) in the age range of $8-12$ years act as a control for the study. All the participants are native Kannada speakers. The manifestation of the CLP speech disorders are performed by three expert speech language pathologists (SLP) who have an experience of a minimum of five years in the field of CLP speech evaluation.

(ii) **Modification of Misarticulated Fricative /s/:** This work involves investigating the acoustic characteristics of misarticulated fricatives for automatic segmentation followed by modification of these errors close to non-CLP /s/. Three types of misarticulated fricative /s/ are analyzed: palatalized articulation, phoneme specific nasal air emission distorted /s/ and glottal stop substituted /s/ in initial and medial position in fricative-vowel-fricative-vowel (FVFV) structure [8]. An automatic segmentation of the misarticulated fricative /s/ is performed using the onset of glottal activity region as an anchoring point. Within 150 ms of the onset point of glottal activity region, the fricative evidence is investigated using band energy ratio and spectral tilt feature. If fricative evidence is detected, then the category of fricative error is determined using the features namely, spectral centroid, dominant spectral centroid, and maximum normalized spectral slope. The detected misarticulated fricative is further subject to modification by applying spectral energy compression followed by spectral tilt modification. On the other hand, if no frication is found, then the search interval is analyzed using short-time energy and abrupt transition characteristics. If the region is identified as silent with steep transitions, it is considered glottal stop substituted /s/. Therefore, the glottal stop substituted /s/ is modified by inserting artificially synthesized /s/.

(iii) **Event-Based Transformation of Misarticulated Stops:** The presence of articulation error in CLP speech degrades the speech intelligibility severely, specifically the stop consonants. Accordingly, this work focuses on the modification of articulation errors, namely, glottal, palatal, and velar stop substitutions produced for the unvoiced stops (/k/, /t/ and /T/). Stop consonants are characterized by dynamically varying spectro-temporal characteristics. Hence, for the modification of stop, the spectral transformation should specifically represent the dynamic characteristics rather than the mixture of different spectral components present in the utterance. Therefore, in this work, an event-based approach is proposed for the spectral transformation. First, automatic detection of burst onset and vowel onset events are carried out. Having detected the acoustic events, the region from burst onset to 20 ms transition followed by vowel onset is segmented. The segmented regions of the source (CLP) and target (non-CLP) speech are used for learning the transformation matrix, which is optimized using nonnegative matrix factorization method. The optimized transformation matrix is further used to modify the articulation errors.

(iv) **Modification of hypernasal vowels using temporal and spectral processing:** The nasalization of vowel reduces the clarity of speech, thus making the speech less intelligible [2, 235]. From the acoustic analysis of hypernasal speech, it is observed that the spectral characteristics of vowels (/a/,/i/, and /u/) are deviated due to the introduction of additional formant and anti-formant pairs in the spectrum, broadening of formant bandwidths and spectral flattening [243]. The hypernasal speech modification is carried out using children speech data. The high-pitch speech affects the lowest formant and hypernasality introduces additional nasal formant in the lower frequency region. This necessitates the accurate representation of the spectral envelope of high-pitch hypernasal speech. Therefore, speech parameterization is performed using extended weighted linear prediction (XLP) method. The XLP coefficient cepstrum from both the source and target speakers are used to build the conversion function for training. The transformation is achieved using GMM based spectral conversion function derived from the source (hypernasal) and target (non-CLP) speakers probabilistic model. It is also noted that, because of the deviated spectral characteristics, the obtained residual signal consists of scaled and delayed versions (interfering components) of the original speech [255]. While synthesis, the interfering signal components in the residual signal may sometimes introduce unnatural spectral changes which are perceived as distortion in the enhanced speech. Therefore, modification of residual and spectral characteristics are expected to result in intelligible speech. A fine weight function is used for de-emphasizing the interfering signal components of the XLP residual signal. The hypernasal speech characteristics are modified while preserving the fundamental frequency of the source speaker.

(v) **Combined framework for the enhancement of an entire word-level intelligibility:** The word-level intelligibility is attempted by combining the specific phoneme modification techniques discussed in the previous chapters. A comparison study is also done to observe whether the transformation method exploited in the above chapters can improve the entire word-level intelligibility. When the transformation method is used to modify the word as a whole, it is observed that the speech sound gets muffled. Therefore, different approaches are exploited to achieve a good quality enhanced speech. Several issues exist in performing the entire word-level intelligibility because many times, both articulation error and hypernasality are observed in the same word. It is challenging to detect such misarticulations in an unsupervised method.

Hence, with certain assumptions and prior knowledge, the enhancement task is carried out.

8.2 Contributions of the thesis

The notable contributions of the thesis are stated as follows:

- A database is developed for studying the CLP speech characteristics. Database comprised of nonsensical words, vowel phonations, meaningful words, and short phrases. However, only some nonsensical words and vowel phonations are used in this thesis.
- Two primary factors of the CLP speech degradation are studied, namely articulation error (misarticulation) and hypernasality.
- Misarticulated fricative /s/ is first studied and modified as it is observed as one of the frequently occurring errors in the database and findings from various studies also support the same.
- Misarticulated unvoiced stops /k/, /t/, and /T/ are analyzed and modified using NMF based spectral conversion.
- Nasalized vowels /a/, /i/, and /u/ are modified using temporal processing and GMM based spectral conversion.
- Finally, phoneme specific modification techniques are combined to achieve entire word-level intelligibility.

8.3 Directions for future work

Based on the studies carried out in the thesis, a few feasible future directions are stated as follows:

(i) **Sub-band based de-nasalization of hypernasal speech:** Hypernasal speech is characterized by spectral deviations with the presence of nasal formants in different frequency locations, consistently in the lower frequency region. Specifically, the lower frequency band consistently exhibit the nasalization characteristics. Therefore processing the spectrum in different frequency bands may result in effective modification of hypernasal speech. Hence, the significance of sub-band based processing of high-pitch hypernasal speech is worth exploring.

(ii) **Exploring CLP speech enhancement for other distorted phonemes:** In the current thesis, we have illustrated CLP speech enhancement using some nonsensical words only in the clinical settings. However, in CLP speech there are other phonemes as well which are distorted due to the articulatory impairment. For instance, phoneme specific nasal air emission distorts

other fricatives and affricates, namely, /z/, /sh/, /zh/, /ch/, and /j/, respectively. In certain cases, nasalization of oral consonants are observed, where the voiced stops are substituted by their nasal cognates, such as, /b/ → /m/, /d/ → /n/, and /g/ → /ng/. Another observation is that when the fricative /s/ is combined with unvoiced stops /p/, /t/, and /k/. The stops are substituted by their voiced cognates, for example, /spell/ → /bell/, /stop/ → /dop/, and /skate/ → /gate/. This necessitates the inclusion of additional CLP speech distortions and the respective phonemes on which they occur frequently.

(iii) **Enhancement of CLP speech by minimizing its difference with the non-CLP speech using iterative approach:** Most of the speech enhancement techniques perform the study based on the acoustic characteristics of clean speech (in the case of non-CLP speaker's degraded speech enhancement) or non-CLP speech template (in the case of disordered speech enhancement). CLP speech enhancement in the current thesis is also studied in the similar fashion. However, the presented approaches were performed on one-shot experiment. The different studies in the thesis showed significant improvement. However, there is scope for further improvement. For this we can explore the study in an iterative manner and observe the outcome, whether we get any improvement in the performance or not?

(iv) **CLP speech enhancement using neural networks:** Recent speech enhancement literatures have shown that neural network based methods become the mainstream strategy for speech modification task. It results in improved speech quality and intelligibility compared to the traditional voice conversion (VC) methods (GMM based VC and NMF based VC) used for pathological speech enhancement [1, 53, 269]. However, the neural network based methods always need a large amount of parallel data to train the system for improving the generalization of the network.

Collecting a large amount of pathological speech data such as CLP speech and creating a parallel corpus is challenging. This projects non-parallel VC methods more suitable for utilizing the benefits of neural network for CLP speech enhancement. To address this issue, exploration of cycle-consistent generative adversarial networks (CycleGAN) based technique would be a good strategy because this approach achieved excellent performance for image-to-image translation, music style transfer, singing voice separation, VC, and noise robust ASR tasks [29] without using parallel data. The CycleGAN is one of the state-of-the-art non-parallel VC methods that

has shown its effectiveness for various applications [144, 148, 149].

Considering CycleGAN to improve the intelligibility of the CLP speech, an initial investigation was carried out in [270]. From the experimental evaluation results, it is observed that the speech signals obtained using CycleGAN method showed an improved performance accuracy compared to original distorted CLP speech. However, it is also observed that the modified CLP speech quality achieves a relatively lower MOS compared to that of the non-CLP speech. This implies that the mapping parameters used during conversion deserves future exploration. A much more improvement in both the quality and intelligibility is desired to make a more robust system in order to bring it close to the target (non-CLP) speech and deal with the realistic environment. Additionally, a thorough experiment must be carried out using nonsensical words, meaningful words, and sentences for handling the spoken-input systems by the CLP speakers.

(v) **Exploring data augmentation approach to improve the ASR performance for CLP speech:** The articulatory impairment in CLP speakers interrupt them from effective use of speech based applications. It is because the system of such devices are trained using non-CLP speakers' speech. Training a robust recognition system specifically for the CLP speakers is challenging because of scarcity of data. Hence, the ability of generative models in learning and mimicking data distributions can be explored to serve the purpose. With generative models, sufficient amount of synthetic speech samples can be generated for augmentation of training data and observe the system performances. Additionally, different data augmentation approaches can also be explored for improving the speech recognition of CLP speech. This work is submitted in APSIPA 2021.

(vi) **CLP speech enhancement for improving ASR performance:** In the thesis, semi-supervised approach is followed to illustrate an improvement in the performance of CLP speech recognition. The approach could be further improved if the speech enhancement techniques are applied for the CLP speech distortions based on the output of the speech recognition system. Then, we could analyze the recognition performance of the enhanced CLP speech, whether it improves or not?

Bibliography

[1] A. B. Kain, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743–759, 2007.

[2] A. W. Kummer, *Cleft Palate & Craniofacial Anomalies: Effects on Speech and Resonance*. Nelson Education, 2013.

[3] S. J. Peterson-Falzone, M. A. Hardin-Jones, and M. P. Karnell, *Cleft Palate Speech*. Mosby St. Louis, 2001.

[4] P. Grunwell and D. Sell, "Speech and cleft palate/velopharyngeal anomalies," *Management of Cleft Lip and Palate. London: Whurr*, 2001.

[5] M. Scipioni, M. Gerosa, D. Giuliani, E. Nöth, and A. Maier, "Intelligibility assessment in children with cleft lip and palate in Italian and German," in *Proceedings of Interspeech*, 2009, pp. 967–970.

[6] A. Maier, C. Hacker, E. Noth, E. Nkenke, T. Haderlein, F. Rosanowski, and M. Schuster, "Intelligibility of children with cleft lip and palate: Evaluation by speech recognition techniques," in *Proceedings of 18th International Conference on Pattern Recognition (ICPR)*, vol. 4, 2006, pp. 274–277.

[7] B. Hutters and K. Brøndsted, "Strategies in cleft palate speech–with special reference to danish," *Cleft Palate Journal*, vol. 24, no. 2, pp. 126–136, 1987.

[8] 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.

[9] V. C. Mathad and S. R. M. Prasanna, "Vowel onset point based screening of misarticulated stops in cleft lip and palate speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 450–460, 2019.

[10] S. Kalita, K. S. Girish, S. R. Mahadeva Prasanna, and S. Dandapat, "Objective assessment of cleft lip and palate speech intelligibility using articulation and hypernasality measures," *The Journal of the Acoustical Society of America*, vol. 146, no. 2, pp. 1164–1175, 2019.

[11] C. P. T. Whitehill and J. C. Chun, "Intelligibility and acceptability in speakers with cleft palate," in *Investigations in Clinical Phonetics and Linguistics*. Psychology Press, 2012, pp. 421–432.

[12] J. Maegawa, R. K. Sells, and D. J. David, "Speech changes after maxillary advancement in 40 cleft lip and palate patients." *The Journal of Craniofacial Surgery*, vol. 9, no. 2, pp. 177–82, 1998.

[13] A. R. M. Bibars, F. S. Alfwaress, A. A.-H. Hamasha, Z. A. Al-Hourani, and K. Almhdawi, "Prosthodontic rehabilitation of arabic speaking individuals with velopharyngeal incompetence: a preliminary study," *The Open Dentistry Journal*, vol. 11, p. 436, 2017.

[14] A. D. Bagnall and D. J. David, "Speech results of cleft palate surgery: two methods of assessment," *British Journal of Plastic Surgery*, vol. 41, no. 5, pp. 488–495, 1988.

[15] J. L. Perry, "Studying the velopharyngeal mechanism through 3d computer reconstructions based on magnetic resonance imaging," *Journal of Oral and Maxillofacial Surgery*, vol. 64, no. 9, pp. 88–89, 2006.

[16] T. Watterson, M. Mancini, T. U. Brancamp, and K. E. Lewis, "Relationship between the perception of hypernasality and social judgments in school-aged children," *The Cleft Palate-Craniofacial Journal*, vol. 50, no. 4, pp. 498–502, 2013.

[17] M. Copeland, "The effects of very early palatal repair on speech," *British Journal of Plastic Surgery*, vol. 43, no. 6, pp. 676–682, 1990.

[18] W. Moore and R. K. Sommers, "Phonetic contexts: Their effects on perceived intelligibility in cleft-palate speakers," *Folia Phoniatrica et Logopaedica*, vol. 27, no. 6, pp. 410–422, 1975.

[19] T. L. Whitehill, "Assessing intelligibility in speakers with cleft palate: a critical review of the literature," *The Cleft Palate-Craniofacial Journal*, vol. 39, no. 1, pp. 50–58, 2002.

[20] J. S. Han, "Percentage of correct consonants, speech intelligibility, and speech acceptability in children with cleft palate," *Communication Sciences & Disorders*, vol. 14, no. 2, pp. 183–199, 2009.

[21] D. J. Zajac, C. Plante, A. Lloyd, and K. L. Haley, "Reliability and validity of a computer-mediated, single-word intelligibility test: Preliminary findings for children with repaired cleft lip and palate," *The Cleft Palate-Craniofacial Journal*, vol. 48, no. 5, pp. 538–549, 2011.

[22] P. Landis and P. T. T. Cuc, "Articulation patterns and speech intelligibility of 54 vietnamese children with unoperated oral clefts: clinical observations and impressions," *The Cleft Palate Journal*, vol. 12, no. 2, pp. 234–243, 1975.

[23] J. Subtelny, R. Van Hattum, and B. Myers, "Ratings and measures of cleft palate speech." *The Cleft Palate Journal*, vol. 9, no. 1, p. 18, 1972.

[24] B. J. McWilliams, "Some factors in the intelligibility of cleft-palate speech," *Journal of Speech and Hearing Disorders*, vol. 19, no. 4, pp. 524–527, 1954.

[25] S. Kalita, P. Mariswamy, A. Abraham, K. S. Girish, S. R. M. Prasanna, and S. Dandapat, "Relative contribution of hypernasality, consonant production errors, and voice disorders on the intelligibility of cleft lip and palate speech," in *Workshop on Speech Processing for Voice, Speech and Hearing Disorders (WSPD)*, 2018.

[26] A. Lohmander and M. Olsson, "Methodology for perceptual assessment of speech in patients with cleft palate: a critical review of the literature," *The Cleft Palate-Craniofacial Journal*, vol. 41, no. 1, pp. 64–70, 2004.

[27] J. L. Locke, "The inference of speech perception in the phonologically disordered child. part ii: Some clinically novel procedures, their use, some findings," *Journal of Speech and Hearing Disorders*, vol. 45, no. 4, pp. 445–468, 1980.

[28] S. Strömbergsson, "The/k/s, the/t/s, and the inbetweens: Novel approaches to examining the perceptual consequences of misarticulated speech," Ph.D. dissertation, KTH Royal Institute of Technology, 2014.

[29] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1826–1838, 2020.

[30] 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.

[31] M. Vucovich, R. R. Hallac, A. A. Kane, J. Cook, C. V. Slot, and J. R. Seaward, "Automated cleft speech evaluation using speech recognition," *Journal of Cranio-Maxillofacial Surgery*, vol. 45, no. 8, pp. 1268–1271, 2017.

[32] F. Ballati, F. Corno, and L. De Russis, "Assessing virtual assistant capabilities with Italian dysarthric speech," in *Proceedings of International ACM SIGACCESS Conference on Computers and Accessibility (ASSET)*, 2018, pp. 93–101.

[33] ——, "Hey Siri, do you understand me?: Virtual assistants and dysarthria." in *Proceedings of International Conference on Intelligent Environments*, 2018, pp. 557–566.

[34] A. Pradhan, K. Mehta, and L. Findlater, "Accessibility came by accident: use of voice-controlled intelligent personal assistants by people with disabilities," in *Proceedings of CHI Conference on Human Factors in Computing Systems*, 2018, p. 459.

[35] C. Yu, R. E. Zezario, S.-S. Wang, J. Sherman, Y.-Y. Hsieh, X. Lu, H.-M. Wang, and Y. Tsao, "Speech enhancement based on denoising autoencoder with multi-branched encoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2756–2769, 2020.

[36] K. Hermus, P. Wambacq *et al.*, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 045821, 2006.

[37] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 700–708, 2003.

[38] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, 2007.

[39] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Transactions on Speech and Audio processing*, vol. 6, no. 5, pp. 445–455, 1998.

[40] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 253–256, 2013.

[41] Y. Zhou, H. Zhao, L. Shang, and T. Liu, "Immune K-SVD algorithm for dictionary learning in speech denoising," *Neurocomputing*, vol. 137, pp. 223–233, 2014.

[42] T. T. Vu, B. Bigot, and E. S. Chng, "Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 499–503.

[43] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 277–282, 1976.

[44] B. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 1, pp. 22–32, 1969.

[45] T. F. Quatieri and R. J. McAulay, "Peak-to-RMS reduction of speech based on a sinusoidal model," *IEEE Transactions on signal processing*, vol. 39, no. 2, pp. 273–288, 1991.

[46] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 1, 2006, pp. I493–I496.

[47] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proceedings of Interspeech*, 2011, pp. 1837–1840.

[48] B. Langner and A. W. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2005, pp. I265–I268.

[49] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[50] J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *The Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 362–378, 2004.

[51] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[52] A. Prakash, M. R. Reddy, and H. A. Murthy, "Improvement of continuous dysarthric speech quality," in *Proceedings of Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016, pp. 43–49.

[53] S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 11, pp. 2584–2594, 2017.

[54] N. Bi and Y. Qi, "Application of speech conversion to alaryngeal speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 2, pp. 97–105, 1997.

[55] H. Liu, Q. Zhao, M. Wan, and S. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 5, pp. 865–874, 2006.

[56] K. Xiao, S. Wang, M. Wan, and L. Wu, "Reconstruction of Mandarin electrolaryngeal fricatives with hybrid noise source," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 2, pp. 383–391, 2019.

[57] W. Li, Q. Zhaopeng, F. Yijun, and N. Haijun, "Design and preliminary evaluation of electrolarynx with F0 control based on capacitive touch technology," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, pp. 629–636, 2018.

[58] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[59] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 172–183, 2013.

[60] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[61] Y. Y. Kong and A. Mullangi, "On the development of a frequency-lowering system that enhances place-of-articulation perception," *Speech Communication*, vol. 54, no. 1, pp. 147–160, 2012.

[62] H. Murakami, S. Hara, M. Abe, M. Sato, and S. Minagi, "Naturalness improvement algorithm for reconstructed glossectomy patients speech using spectral differential modification in voice conversion," *Proceedings of Interspeech*, pp. 2464–2468, 2018.

[63] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

[64] A. Rao and L. H. Carney, "Speech enhancement for listeners with hearing loss based on a model for vowel coding in the auditory midbrain," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 7, pp. 2081–2091, 2014.

[65] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029–3038, 2013.

[66] T. Nguyen, R. Garrett, A. Downing, L. Walker, and D. Hobbs, "An interfacing system that enables speech generating device users to independently access and use a mobile phone," *Technology and Disability*, vol. 20, no. 3, pp. 225–239, 2008.

[67] E. Simion, "Augmentative and alternative communication–support for people with severe speech disorders," *Procedia-Social and Behavioral Sciences*, vol. 128, pp. 77–81, 2014.

[68] P. Kitzing, A. Maier, and V. L. Åhlander, "Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders," *Logopedics Phoniatrics Vocology*, vol. 34, no. 2, pp. 91–96, 2009.

[69] N. Thomas Stonell, A. L. Kotler, H. Leeper, and P. Doyle, "Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition accuracy," *Augmentative and Alternative Communication*, vol. 14, no. 1, pp. 51–56, 1998.

[70] N. Jamal, S. Shanta, F. Mahmud, and M. Shaabani, "Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review," in *Proceedings of AIP*, vol. 1883, no. 1, 2017, p. 020028.

[71] 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.

[72] R. Kataoka, D. J. Zajac, R. Mayo, R. W. Lutz, and D. W. Warren, "The influence of acoustic and perceptual factors on perceived hypernasality in the vowel [i]: A preliminary study," *Folia Phoniatrica et Logopaedica*, vol. 53, no. 4, pp. 198–212, 2001.

[73] A. S. Y. Lee, V. Ciocca, and T. L. Whitehill, "Acoustic correlates of hypernasality," *Clinical Linguistics & Phonetics*, vol. 17, no. 4-5, pp. 259–264, 2003.

[74] S. Ha, H. Sim, M. Zhi, and D. P. Kuehn, "An acoustic study of the temporal characteristics of nasalization in children with and without cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 41, no. 5, pp. 535–543, 2004.

[75] "Speech sound disorders: Articulation and phonological processes," Retrieved April 20, 2015 from, https://www.asha.org/public/speech/disorders/speech-sound-disorders/.

[76] G. Parikh and P. C. Loizou, "The influence of noise on vowel and consonant cues," *The Journal of the Acoustical Society of America*, vol. 118, no. 6, pp. 3874–3888, 2005.

[77] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[78] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 7, p. 354850, 2005.

[79] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[80] V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 764–773, 2006.

[81] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[82] H. Taşmaz and E. Erçelebi, "Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE–STSA estimation in various noise environments," *Digital Signal Processing*, vol. 18, no. 5, pp. 797–812, 2008.

[83] R. Koning, N. Madhu, and J. Wouters, "Ideal time–frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 331–341, 2014.

[84] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1948–1963, 2012.

[85] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[86] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[87] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[88] T. Mellahi and R. Hamdi, "LPC-based formant enhancement method in Kalman filtering for speech enhancement," *AEU-International Journal of Electronics and Communications*, vol. 69, no. 2, pp. 545–554, 2015.

[89] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, "Nonlinear speech enhancement: An overview," in *Progress in Nonlinear Speech Processing*. Springer, 2007, pp. 217–248.

[90] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[91] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[92] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[93] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2013.

[94] ——, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[95] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement." in *Proceedings of Interspeech*, 2016, pp. 3768–3772.

[96] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 006–012.

[97] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[98] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[99] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proceedings of Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 136–140.

[100] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[101] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[102] Y. C. Subakan and P. Smaragdis, "Generative adversarial source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 26–30.

[103] K. Wang, J. Zhang, S. Sun, Y. Wang, F. Xiang, and L. Xie, "Investigating generative adversarial networks based speech dereverberation for robust speech recognition," *arXiv preprint arXiv:1803.10132*, 2018.

[104] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5024–5028.

[105] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[106] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*, 2013, pp. 7092–7096.

[107] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2015.

[108] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[109] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Proceedings of Interspeech*, 2017, pp. 1993–1997.

[110] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis," in *Proceedings of 5th ISCA Workshop on Speech Synthesis*, 2004.

[111] R. Alexander, T. Sorensen, A. Toutios, and S. Narayanan, "A modular architecture for articulatory synthesis from gestural specification," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4458–4471, 2019.

[112] D. R. Hill, C. R. Taube-Schock, and L. Manzara, "Low-level articulatory synthesis: A working text-to-speech solution and a linguistic tool," *The Canadian Journal of Linguistics/La revue canadienne de linguistique*, vol. 62, no. 3, pp. 371–410, 2017.

[113] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal folds," in *Proceedings of 1st International Workshop on Performative Speech and Singing Synthesis*, vol. 370, 2011.

[114] Y. Tabet and M. Boughazi, "Speech synthesis techniques. A survey," in *Proceedings of International Workshop on Systems, Signal Processing and their Applications, WOSSPA*, 2011, pp. 67–70.

[115] K. H. Wee, L. Turicchia, and R. Sarpeshkar, "An articulatory speech-prosthesis system," in *Proceedings of IEEE International Conference on Body Sensor Networks*, 2010, pp. 133–138.

[116] R. E. Donovan, "Trainable speech synthesis," Ph.D. dissertation, University of Cambridge, 1996.

[117] K. Matsui, N. Hara, N. Kobayashi, and H. Hirose, "Enhancement of esophageal speech using formant synthesis," *Acoustical Science and Technology*, vol. 23, no. 2, pp. 69–76, 2002.

[118] W. J. Holmes, J. N. Holmes, and M. W. Judd, "Extension of the bandwidth of the JSRU parallel-formant synthesizer for high quality synthesis of male and female speech," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1990, pp. 313–316.

[119] M. S. Scordilis, "A neuronal formant synthesizer," *Mathematics and Computers in Simulation*, vol. 40, no. 5-6, pp. 615–622, 1996.

[120] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Comparison of formant enhancement methods for HMM-based speech synthesis," in *Proceedings of 7th ISCA Workshop on Speech Synthesis*, 2010, pp. 334–339.

[121] M. Plumpe, A. Acero, H.-W. Hon, and X. Huang, "HMM-based smoothing for concatenative speech synthesis," in *Proceedings of 5th International Conference on Spoken Language Processing*, 1998.

[122] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1996, pp. 373–376.

[123] J. P. Van Santen, R. Sproat, J. Olive, and J. Hirschberg, *Progress in Speech Synthesis*. Springer Science & Business Media, 2013.

[124] K. Kuligowska, P. Kisielewicz, and A. Włodarz, "Speech synthesis systems: disadvantages and limitations," *International Journal of Engineering & Technology*, vol. 7, no. 2.28, pp. 234–239, 2018.

[125] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[126] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[127] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1998, pp. 285–288.

[128] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 2802–2806.

[129] Z. Wu and H. Li, "Voice conversion versus speaker verification: an overview," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[130] O. Turk and L. M. Arslan, "Robust processing techniques for voice conversion," *Computer Speech & Language*, vol. 20, no. 4, pp. 441–467, 2006.

[131] K. Oyamada, H. Kameoka, T. Kaneko, H. Ando, K. Hiramatsu, and K. Kashino, "Non-native speech conversion with consistency-aware recursive network and generative adversarial network," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 182–188.

[132] C. Veaux, J. Yamagishi, and S. King, "Towards personalised synthesised voices for individuals with vocal disabilities: Voice banking and reconstruction," in *Proceedings of the 4th Workshop on Speech and Language Processing for Assistive Technologies*, 2013, pp. 107–111.

[133] K.-S. Lee, "Statistical approach for voice personality transformation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 641–651, 2007.

[134] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," in *Proceedings of Interspeech*, 2008, pp. 1453–1456.

[135] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1988, pp. 655–658.

[136] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[137] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[138] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[139] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 3, pp. 806–817, 2011.

[140] R. Vích and M. Vondra, "Pitch synchronous transform warping in voice conversion," in *Cognitive Behavioural Systems*. Springer, 2012, pp. 280–289.

[141] B. Sisman, M. Zhang, M. Dong, and H. Li, "On the study of generative adversarial networks for cross-lingual voice conversion," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 144–151.

[142] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," in *Proceedings of Interspeech*, 2017, pp. 3364–3368.

[143] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 266–273.

[144] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proceedings of 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100–2104.

[145] Y. Luan, D. Saito, Y. Kashiwagi, N. Minematsu, and K. Hirose, "Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 1745–1748.

[146] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[147] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[148] C. c. Yeh, P. c. Hsu, J. c. Chou, H. y. Lee, and L. s. Lee, "Rhythm-flexible voice conversion without parallel data using cycle-GAN over phoneme posteriorgram sequences," in *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 274–281.

[149] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5279–5283.

[150] K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka, "Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks," in *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 632–639.

[151] G. Kim, H. Lee, B.-K. Kim, S.-H. Oh, and S.-Y. Lee, "Unpaired speech enhancement by acoustic and adversarial supervision for speech recognition," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 159–163, 2018.

[152] Z. Meng, J. Li, Y. Gong, and B.-H. F. Juang, "Cycle-consistent speech enhancement," in *Proceedings of Interspeech*, 2018, pp. 1165–1169.

[153] O. Lachhab, J. Di Martino, E. I. Elhaj, and A. Hammouch, "A preliminary study on improving the recognition of esophageal speech using a hybrid system based on statistical voice conversion," *SpringerPlus*, vol. 4, no. 1, pp. 1–14, 2015.

[154] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Elsevier Health Sciences, 2019.

[155] S. Chandrakala and N. Rajeswari, "Representation learning based speech assistive system for persons with dysarthria," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1510–1517, 2016.

[156] M. Dhanalakshmi, T. M. Celin, T. Nagarajan, and P. Vijayalakshmi, "Speech-input speech-output communication for dysarthric speakers using HMM-based speech recognition and adaptive synthesis system," *Circuits, Systems, and Signal Processing*, vol. 37, no. 2, pp. 674–703, 2018.

[157] J.-P. Hosom, A. B. Kain, T. Mishra, J. P. Van Santen, M. Fried-Oken, and J. Staehely, "Intelligibility of modifications to dysarthric speech," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2003, pp. I924–I927.

[158] C. Shilpa, V. Swathi, V. Karjigi, K. Pavithra, and S. Sultana, "Landmark based modification to correct distortions in dysarthric speech," in *Proceedings of 22nd National Conference on Communication (NCC)*, 2016, pp. 1–6.

[159] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 8037–8040.

[160] ——, "Consonant enhancement for articulation disorders based on non-negative matrix factorization," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2012, pp. 1–4.

[161] J. C. Kim, H. Rao, and M. A. Clements, "Speech intelligibility estimation using multi-resolution spectral features for speakers undergoing cancer treatment," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. EL315–EL321, 2014.

[162] K. Tanaka, S. Hara, M. Abe, and S. Minagi, "Enhancing a glossectomy patient's speech via GMM-based voice conversion," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2016, pp. 1–4.

[163] J.-P. Laaksonen, I. J. Loewen, J. Wolfaardt, J. Rieger, H. Seikaly, and J. Harris, "Speech after tongue reconstruction and use of a palatal augmentation prosthesis: An acoustic case study." *Canadian Journal of Speech-Language Pathology & Audiology*, vol. 33, no. 4, 2009.

[164] V. d. Carvalho and L. U. Sennes, "Speech and swallowing data in individual patients who underwent glossectomy after prosthetic rehabilitation," *International Journal of Dentistry*, vol. 2016, pp. 1–11.

[165] K. Mády, R. Sader, P. Hoole, A. Zimmermann, and H.-H. Horch, "Speech evaluation and swallowing ability after intra-oral cancer," *Clinical Linguistics & Phonetics*, vol. 17, no. 4-5, pp. 411–420, 2003.

[166] L. W. Chen, H. Y. Lee, and Y. Tsao, "Generative adversarial networks for unpaired voice transformation on impaired speech," in *Proceedings of Interspeech*, 2019, pp. 719–723.

[167] G. H. Saunders and J. M. Kates, "Speech intelligibility enhancement using hearing-aid array processing," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1827–1837, 1997.

[168] Y.-T. Liu, Y. Tsao, and R. Y. Chang, "Nonnegative matrix factorization-based frequency lowering technology for Mandarin-speaking hearing aid users," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 5905–5909.

[169] C. K. A. Reddy, N. Shankar, G. S. Bhat, R. Charan, and I. Panahi, "An individualized super-Gaussian single microphone speech enhancement for hearing aid users with smartphone as an assistive device," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1601–1605, 2017.

[170] Y.-H. Lai and W.-Z. Zheng, "Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users," *Biomedical Signal Processing and Control*, vol. 48, pp. 35–45, 2019.

[171] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proceedings of Interspeech*, 2019, pp. 4115–4119.

[172] O. Watts, J. Yamagishi, S. King, and K. Berkling, "Synthesis of child speech with HMM adaptation and voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1005–1016, 2009.

[173] "Bruel and kjær," Retrieved July 15, 2016 from, https://www.bksv.com/en.

[174] P. Boersma and V. Van Heuven, "Speak and unSpeak with PRAAT," *Glot International*, vol. 5, no. 9-10, pp. 341–347, 2001.

[175] D. Liljequist, B. Elfving, and K. S. Roaldsen, "Intraclass correlation–A discussion and demonstration of basic features," *PloS One*, vol. 14, no. 7, 2019.

[176] P. N. Sudro and S. R. M. Prasanna, "Modification of misarticulated fricative /s/ in cleft lip and palate speech," *Biomedical Signal Processing and Control*, vol. 67, p. 102088, 2021.

[177] S. Kalita, P. N. Sudro, S. R. M. Prasanna, and S. Dandapat, "Nasal air emission in sibilant fricatives of cleft lip and palate speech," in *Proceedings of Interspeech*, 2019, pp. 4544–4548.

[178] P. N. Sudro, S. Kalita, and S. R. M. Prasanna, "Processing transition regions of glottal stop substituted /s/ for intelligibility enhancement of cleft palate speech," in *Proceedings of Interspeech*, 2018, pp. 1536–1540.

[179] P. N. Sudro and S. R. M. Prasanna, "Intelligibility enhancement of alveolar fricative in cleft lip and palate speech," in *Proceedings of 14th IEEE India Council International Conference (INDICON)*, 2017, pp. 1–6.

[180] K. N. Stevens, "Airflow and turbulence noise for fricative and stop consonants: Static considerations," *The Journal of the Acoustical Society of America*, vol. 50, no. 4B, pp. 1180–1192, 1971.

[181] K. Iskarous, C. H. Shadle, and M. I. Proctor, "Articulatory–acoustic kinematics: The production of American English /s/," *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 944–954, 2011.

[182] L. F. Wilde, "Analysis and synthesis of fricative consonants," Ph.D. dissertation, Massachusetts Institute of Technology, 1995.

[183] R. Romeo, V. Hazan, and M. Pettinato, "Developmental and gender-related trends of intra-talker variability in consonant production," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3781–3792, 2013.

[184] B. Sharma and S. R. M. Prasanna, "Enhancement of spectral tilt in synthesized speech," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 382–386, 2017.

[185] J.-W. Lee, J.-Y. Choi, and H.-G. Kang, "Classification of fricatives using feature extrapolation of acoustic-phonetic features in telephone speech," in *Proceedings of Interspeech*, 2011, pp. 1261–1264.

[186] M. Hodge and C. L. Gotzke, "Preliminary results of an intelligibility measure for English-speaking children with cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 44, no. 2, pp. 163–174, 2007.

[187] M. Tezuka, Y. Ogata, K. Matsunaga, T. Mitsuyasu, S. Hasegawa, and N. Nakamura, "Perceptual and videofluoroscopic analyses of relation between backed articulation and velopharyngeal closure following cleft palate repair," *Oral Science International*, vol. 11, no. 2, pp. 60–67, 2014.

[188] T. Arai, K. Okazaki, S. Imatomi, and Y. Yoshida, "Analysis for palatalized articulation of [s] sounds using synthetic speech," in *EUROSPEECH*, 1995, pp. 1–4.

[189] D. J. Zajac, R. Mayo, R. Kataoka, and J. Y. Kuo, "Aerodynamic and acoustic characteristics of a speaker with turbulent nasal emission: A case report," *The Cleft Palate-Craniofacial Journal*, vol. 33, no. 5, pp. 440–444, 1996.

[190] A. L. Baylis, B. Munson, and K. T. Moller, "Perceptions of audible nasal emission in speakers with cleft palate: a comparative study of listener judgments," *The Cleft Palate-Praniofacial Journal*, vol. 48, no. 4, pp. 399–411, 2011.

[191] R. N. Ohde, D. J. Sharf, and P. F. Jacobson, "Phonetic analysis of normal and abnormal speech," in *The Journal of Acoustical Society of America*, 1992, p. 3452.

[192] D. P. Kuehn and L. J. Henne, "Speech evaluation and treatment for patients with cleft palate," *American Journal of Speech-Language Pathology*, 2003.

[193] L. He, J. Zhang, Q. Liu, J. Zhang, H. Yin, and M. Lech, "Automatic detection of glottal stop in cleft palate speech," *Biomedical Signal Processing and Control*, vol. 39, pp. 230–236, 2018.

[194] 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.

[195] N. Kido, M. Kawano, F. Tanokuchi, Y. Fujiwara, I. Honjo, and H. Kojima, "Glottal stop in cleft palate speech," *Studia Phonologica*, vol. 26, pp. 34–41, 1992.

[196] N. J. Lass, *Speech and Language: Advances in Basic Research and Practice*. Academic Press, 2014, vol. 5.

[197] K. S. R. 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.

[198] V. C. M., S. R. M. Prasanna, A. K. Abraham, P. M, and G. K. S, "Detection of glottal activity errors in production of stop consonants in children with cleft lip and palate," in *Proceedings of Interspeech*, 2018, pp. 382–386.

[199] A. M. Abdelatty Ali, J. Van der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of fricatives," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2217–2235, 2001.

[200] K. Nataraj, P. C. Pandey, and H. Dasgupta, "Estimation of place of articulation of fricatives from spectral characteristics for speech training," *Proceedings of Interspeech*, pp. 339–343, 2017.

[201] E. F. Beach and C. Kitamura, "Modified spectral tilt affects older, but not younger, infants' native-language fricative discrimination," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 2, pp. 658–667, 2011.

[202] K. Singh and N. Tiwari, "The structure of Hindi stop consonants," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. 3633–3642, 2016.

[203] P. G. Stelmachowicz, K. Nishi, S. Choi, D. E. Lewis, B. M. Hoover, D. Dierking, and A. Lotto, "Effects of stimulus bandwidth on the imitation of English fricatives by normal-hearing children," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 5, pp. 1369–1380, 2008.

[204] R. K. Shosted, "The aeroacoustics of nasalized fricatives," Ph.D. dissertation, University of California, 2006.

[205] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The KALDI speech recognition toolkit," in *Proceedings of Workshop on Automatic Speech Recognition and Understanding*, 2011.

[206] D. M. Shiller, M. Sato, V. L. Gracco, and S. R. Baum, "Perceptual recalibration of speech sounds following speech motor learning," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1103–1113, 2009.

[207] P. N. Sudro, C. Vikram, and S. R. M. Prasanna, "Event-based transformation of misarticulated stops in cleft lip and palate speech," *Circuits, Systems, and Signal Processing*, pp. 1–25, 2021.

[208] P. N. Sudro, "Intelligibility enhancement of cleft lip and palate speech," pp. 1–3, 2019.

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

[210] T. Ananthapadmanabha, A. Prathosh, and A. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 460–471, 2014.

[211] A. Prathosh, A. Ramakrishnan, and T. Ananthapadmanabha, "Estimation of voice-onset time in continuous speech using temporal measures," *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. EL122–EL128, 2014.

[212] L. Santelmann, J. Sussman, and K. Chapman, "Perception of middorsum palatal stops from the speech of three children with repaired cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 36, no. 3, pp. 233–242, 1999.

[213] F. E. Gibbon, "Abnormal patterns of tongue-palate contact in the speech of individuals with cleft palate," *Clinical Linguistics & Phonetics*, vol. 18, no. 4-5, pp. 285–311, 2004.

[214] F. E. Gibbon, L. Ellis, and L. Crampin, "Articulatory placement for/t/,/d/,/k/, and /g/ targets in school age children with speech disorders associated with cleft palate," *Clinical Linguistics & Phonetics*, vol. 18, no. 6-8, pp. 391–404, 2004.

[215] A. Harding and P. Grunwell, "Characteristics of cleft palate speech," *International Journal of Language & Communication Disorders*, vol. 31, no. 4, pp. 331–357, 1996.

[216] M. Eshghi, D. J. Zajac, M. Bijankhan, and M. Shirazi, "Spectral analysis of word-initial alveolar and velar plosives produced by Iranian children with cleft lip and palate," *Clinical Linguistics & Phonetics*, vol. 27, no. 3, pp. 213–219, 2013.

[217] Y. Xiao, Y. Feng, Q. Zhao, L. Ma, J. Qian, and Y. Yan, "Acoustic analysis and detection of glottal stops substituted for alveolar stops in cleft palate speech," *Shengxue Xuebao/Acta Acustica*, vol. 40, no. 2, pp. 285–293, 2015.

[218] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 769–773, 1955.

[219] D. Kewley-Port, "Measurement of formant transitions in naturally produced stop consonant–vowel syllables," *The Journal of the Acoustical Society of America*, vol. 72, no. 2, pp. 379–389, 1982.

[220] B. J. Philips and R. D. Kent, "Acoustic–phonetic descriptions of speech production in speakers with cleft palate and other velopharyngeal disorders," in *Speech and Language*. Elsevier, 1984, vol. 11, pp. 113–168.

[221] P. Jain and R. B. Pachori, "Event-based method for instantaneous fundamental frequency estimation from voiced speech based on eigenvalue decomposition of the Hankel matrix," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1467–1482, 2014.

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

[223] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[224] Z. Wu, E. S. Chng, and H. Li, "Joint nonnegative matrix factorization for exemplar-based voice conversion," in *Proceedings of Interspeech*, 2014, pp. 2509–2513.

[225] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, 1996.

[226] A. Prathosh, T. Ananthapadmanabha, and A. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2471–2480, 2013.

[227] P. K. Ghosh and S. S. Narayanan, "Closure duration analysis of incomplete stop consonants due to stop-stop interaction," *The Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. EL1–EL7, 2009.

[228] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *The Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, 1975.

[229] V. Karjigi and P. Rao, "Classification of place of articulation in unvoiced stops with spectro-temporal surface modeling," *Speech Communication*, vol. 54, no. 10, pp. 1104–1120, 2012.

[230] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[231] Y. Zhao, M. Kuruvilla-Dugdale, and M. Song, "Structured sparse spectral transforms and structural measures for voice conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 12, pp. 2267–2276, 2018.

[232] K. N. Stevens, *Acoustic Phonetics*. MIT press, 2000, vol. 30.

[233] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The Mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1–18, 2000.

[234] P. N. Sudro and S. R. M. Prasanna, "Enhancement of cleft palate speech using temporal and spectral processing," *Speech Communication*, vol. 123, pp. 70–82, 2020.

[235] J. Trost-Cardamone, "Diagnosis of specific cleft palate speech error patterns for planning therapy or physical management needs," *Communicative Disorders Related to Cleft Lip and Palate*, 1997.

[236] M. S. Frederickson, K. L. Chapman, and M. Hardin-Jones, "Conversational skills of children with cleft lip and palate: a replication and extension," *The Cleft Palate-Craniofacial Journal*, vol. 43, no. 2, pp. 179–188, 2006.

[237] J. H. N. Pinto, G. S. Dalben, and M. I. Pegoraro-Krook, "Speech intelligibility of patients with cleft lip and palate after placement of speech prosthesis," *The Cleft Palate-Craniofacial Journal*, vol. 44, no. 6, pp. 635–641, 2007.

[238] 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. 1560–1575, 1985.

[239] G. Fant, *Acoustic Theory of Speech Production*. Walter de Gruyter, 1970, no. 2.

[240] D. A. Cairns, J. H. Hansen, and J. E. Riski, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 1, p. 35, 1996.

[241] 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.

[242] P. Enderby, C. Pickstone, A. John, K. Fryer, A. Cantrell, and D. Papaioannou, "Resource manual for commissioning and planning services for SLCN," *London: Royal College of Speech and Language Therapists*, 2009.

[243] 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.

[244] 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 *Proceedings of Twenty Second National Conference on Communication (NCC)*, 2016, pp. 1–6.

[245] A. K. Dubey, S. M. Prasanna, and S. Dandapat, "Zero time windowing based severity analysis of hypernasal speech," in *Proceedings of IEEE Region 10 Conference (TENCON)*, 2016, pp. 970–974.

[246] J. R. Orozco-Arroyave, S. Murillo-Rendón, A. M. Álvarez-Meza, J. D. Arias-Londoño, E. Delgado-Trejos, J. ú. F. Vargas-Bonilla, and C. G. Castellanos-Domínguez, "Automatic selection of acoustic and non-linear dynamic features in voice signals for hypernasality detection," in *Proceedings of Interspeech*, 2011, pp. 529–532.

[247] S. M. Rendón, J. O. Arroyave, J. V. Bonilla, J. A. Londono, and C. C. Domínguez, "Automatic detection of hypernasality in children," in *Proceedings of International Work-Conference on the Interplay Between Natural and Artificial Computation*, 2011, pp. 167–174.

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

[249] E. Delgado, F. Sepúlveda, S. Röthlisberger, and G. Castellanos, "The rademacher complexity model over acoustic features for improving robustness in hypernasal speech detection," *Computers and Simulation in Modern Science*, vol. 5, pp. 130–135, 2011.

[250] 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.

[251] A. K. Dubey, A. Tripathi, S. R. M. Prasanna, and S. Dandapat, "Detection of hypernasality based on vowel space area," *The Journal of the Acoustical Society of America*, vol. 143, no. 5, pp. EL412–EL417, 2018.

[252] K. Nikitha, S. Kalita, C. Vikram, M. Pushpavathi, and S. M. Prasanna, "Hypernasality severity analysis in cleft lip and palate speech using vowel space area." in *Proceedings of Interspeech*, 2017, pp. 1829–1833.

[253] 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.

[254] 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 Letter*, vol. 21, no. 10, pp. 1298–1301, 2014.

[255] 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.

[256] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[257] R. Sinha and S. Ghai, "On the use of pitch normalization for improving children's speech recognition," in *Proceedings of Interspeech*, 2009, pp. 568–571.

[258] A. K. Dubey, S. R. M. Prasanna, and S. Dandapat, "Pitch-adaptive front-end feature for hypernasality detection," in *Proceedings of Interspeech*, 2018, pp. 372–376.

[259] C. Ma, Y. Kamp, and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 1, pp. 69–81, 1993.

[260] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[261] K. Van Lierde, M. De Bodt, J. Van Borsel, F. Wuyts, and P. Van Cauwenberge, "Effect of cleft type on overall speech intelligibility and resonance," *Folia Phoniatrica et Logopaedica*, vol. 54, no. 3, pp. 158–168, 2002.

[262] 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.

[263] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," in *Proceedings of Interspeech*, 2010, pp. 1477–1480.

[264] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-hall Englewood Cliffs, NJ, 1978, vol. 100.

[265] V. C. Raykar, B. Yegnanarayana, S. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 751–761, 2005.

[266] P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication*, vol. 53, no. 2, pp. 154–174, 2011.

[267] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 6405–6409.

[268] P. N. Sudro, R. Sinha, and S. R. M. Prasanna, "Processing phoneme specific segments for cleft lip and palate speech enhancement," in *Proceedings of 13th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1–6.

[269] P. N. Sudro and S. M. Prasanna, "Modification of devoicing error in cleft lip and palate speech," in *Proceedings of Interspeech*, 2019, pp. 4519–4523.

[270] P. N. Sudro, R. K. Das, R. Sinha, and S. R. M. Prasanna, "Enhancing the intelligibility of cleft lip and palate speech using Cycle-consistent adversarial networks," in *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 720–727.

List of Publications

Journal Publications

- Protima Nomo Sudro, C. M. Vikram , S. R. Mahadeva Prasanna, "Event-Based Transformation of Misarticulated Stops in Cleft Lip and Palate Speech", *Circuits, Systems, and Signal Processing*, 40(8), 2021: 4064-4088.
- Protima Nomo Sudro, S. R. Mahadeva Prasanna,"Modification of misarticulated fricative /s/ in cleft lip and palate speech", *Biomedical Signal Processing and Control* 67 (2021) 102088.
- Protima Nomo Sudro, S. R. Mahadeva Prasanna,"Enhancement of cleft palate speech using temporal and spectral processing", *Speech Communication* 123 (2020): 70-82.

Conference Publications

- Protima Nomo Sudro, " Intelligibility Enhancement of Cleft Lip and Palate Speech", 5^{th} *Doctoral consortium, Interspeech* 2019, Graz, Austria, September 2019.
- Sishir Kalita, Protima Nomo Sudro, S. R. Mahadeva Prasanna, S. Dandapat, "Nasal Air Emission in Sibilant Fricatives of Cleft Lip and Palate Speech", in Proceedings of Interspeech 2019.
- Protima Nomo Sudro, Sishir Kalita, S. R. Mahadeva Prasanna,"Processing Transition Regions of Glottal Stop Substituted /S/ for Intelligibility Enhancement of Cleft Palate Speech", in Proceedings of Interspeech 2018.
- Protima Nomo Sudro, S. R. Mahadeva Prasanna,"Intelligibility Enhancement of Alveolar Fricative in Cleft Lip and Palate Speech", in Proceedings of INDICON 2017.

Other Conference Publications

- Protima Nomo Sudro, Rohan Kumar Das, Rohit Sinha, S. R. Mahadeva Prasanna, "On the importance of data augmentation for cleft lip and palate speech recognition", in Proceedings of APSIPA 2021.
- Protima Nomo Sudro, Rohit Sinha, S. R. Mahadeva Prasanna, " Processing phoneme specific segments for cleft lip and palate speech enhancement" , in Proceedings of APSIPA 2021.
- Protima Nomo Sudro, Rohan Kumar Das, Rohit Sinha, S. R. Mahadeva Prasanna, "Enhancing the intelligibility of cleft lip and palate using cycle-consistent adversarial networks", in Proceedings of SLT, 2020.
- Protima Nomo Sudro, S. R. Mahadeva Prasanna, "Modification of Devoicing Error in Cleft Lip and Palate Speech", in Proceedings of Interspeech 2019.

- Protima Nomo Sudro, Vikram C. M, S. R. Mahadeva Prasanna, "Vowel Onset Point Based Characterization of Velopharyngeal Activity Using Imaging Techniques", in Proceedings of NCC 2017.